THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (INP Toulouse)

Discipline ou spécialité :

Intelligence Artificielle

Présentée et soutenue par :

M. MOHAMMAD GHASEMI HAMED

le jeudi 20 février 2014

Titre :

METHODES NON-PARAMETRIQUES POUR LA PREVISION D'INTERVALLES AVEC HAUT NIVEAU DE CONFIANCE: APPLICATION A LA PREVISION DE TRAJECTOIRES D'AVIONS.

Ecole doctorale :

Mathématiques, Informatique, Télécommunications de Toulouse (MITT)

Unité de recherche :

Institut de Recherche en Informatique de Toulouse (I.R.I.T.)

Directeur(s) de Thèse :

M. NICOLAS DURAND

M. MATHIEU SERRURIER

Rapporteurs :

M. ERIC FERON, GEORGIA INSTITUTE OF TECHNOLOGY

M. EYKE HULLERMEIER, PHILLIPS-UNIVERSITAT MARBURG

Membre(s) du jury :

M. GILLES RICHARD, UNIVERSITE TOULOUSE 3, Président

M. MATHIEU SERRURIER, UNIVERSITE TOULOUSE 3, Membre

M. NICOLAS DURAND, ECOLE NATIONALE DE L'AVIATION CIVILE, Membre

M. SEBASTIEN DESTERCKE, UNIVERSITE DE TECHNOLOGIE DE COMPIEGNE, Membre

"There is no wealth like knowledge, no poverty like ignorance."

Ali ibn Abitaleb

To Amirul Momenin Ali ibn Abitaleb.

Abstract

Ground-based aircraft trajectory prediction is a critical issue for air traffic management. A safe and efficient prediction is a prerequisite for the implementation of automated tools that detect and solve conflicts between trajectories. In this scope, this work proposes two non-parametric interval prediction methods in the regression context. These methods are designed to predict intervals that contain at least a desired proportion of the conditional distribution of the response value (referred to predictive intervals). Firstly, we consider the problem of the estimation of a probability distribution with a small sample size. Based on the probabilistic interpretation of the possibility theory, we describe possibility distributions that encode different kinds of statistical interval. Then, we propose a statistical test to verify the reliability of an interval prediction model. We also introduce two measures for comparing different interval prediction models giving intervals that have different sizes and coverage. Starting from our work on statistical intervals (and the associated possibility distribution), we present a pair of methods to find two-sided predictive intervals for non-parametric least squares regression without the non-biased prediction and the error homoscedasticity assumptions. Our predictive intervals are built by using tolerance intervals on prediction errors in the query point neighborhood. The query point neighborhood is obtained with a fixed or variable size neighborhood selection method. We finally obtain a method that finds in most cases the smallest reliable predictive interval model of a dataset. The proposed interval prediction methods are compared with other well-known interval prediction methods both at the theoretical and the practical level. An evaluation is performed with nine benchmark datasets. They are tested on their reliability, efficiency, precision and tightness of their obtained envelope. These experiments show that our methods are more reliable, effective and precise than their competitors. The final chapter describes the application of our method to an aircraft trajectory prediction problem in the climb phase and we compare the results with those obtained with the state of the art algorithms and with physical models.

Résumé

La prédiction de trajectoires d'avions à partir des données disponibles au sol est un problème critique pour le contrôle aérien. Une prédiction fiable et efficace est un prérequis pour l'implémentation d'outils automatiques pour la détection et la résolution de conflits entre les trajectoires. Dans ce contexte, nous proposons de nouvelles méthodes non paramétriques pour la prédiction d'intervalle contenant une proportion attendue des données avec un haut niveau de confiance. Dans un premier temps, nous traitons le problème de l'estimation d'une distribution de probabilité à partir d'un petit échantillon. En considérant l'interprétation des distributions de possibilité comme une famille de distributions de probabilité, nous décrivons un ensemble de distributions de possibilité qui résument différents types d'intervalles statistiques. Ensuite, nous proposons un cadre de travail pour vérifier si un modèle, construit à partir de données, respecte les propriétés de recouvrement requises par les intervalles de prédiction. Nous introduisons aussi deux mesures pour comparer des modèles de prédiction d'intervalle qui ont des tailles moyennes et des taux de recouvrement différents. À partir de nos travaux sur les intervalles statistiques (et leurs distributions de possibilité associés), nous présentons une nouvelle méthode pour induire des intervalles de prédictions bornés pour des méthodes de régression des moindres carrés non paramétriques sans assumer que la prédiction est non biaisée et que les erreurs sont homoscédastiques. Nos intervalles de prédiction sont construits en utilisant des intervalles de tolérances sur les erreurs dans le voisinage du point à prédire. Pour cela, nous décrivons une méthode de sélection de voisinage à taille fixe ou de voisinage à taille variable dépendant de la quantité d'informations autour du point. Nous obtenons un algorithme qui induit, dans la majorité des cas, les intervalles de prédiction fiables les plus petits possibles. Les méthodes que nous proposons sont comparées avec les méthodes les plus connues au niveau théorique et au niveau pratique. Une évaluation est effectuée sur neuf bases de données. La taille, l'efficacité, la fiabilité et la précision des intervalles prédits sont comparés. Ces expérimentations montrent que nos approches sont significativement plus précises et fiables que les autres. Enfin nous appliquons nos méthodes au problème de la prédiction de trajectoires d'avions et nous comparons les résultats avec ceux des méthodes classiques et des modèles physiques.

Acknowledgments

First and foremost praises and thanks to the God, the Almighty, for His showers of blessings throughout my work to complete this research successfully.

I would like to express my deepest appreciation and thanks to my advisor Dr. Mathieu Serrurier, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I would like to express my special appreciation to my thesis director, Pr. Nicolas Durand whose expertise, understanding, and patience, added considerably to my graduate experience.

My sincere thanks goes to Dr. Steve Lawford who generously accepted to correct this work.

I would also like to thank my committee members, Pr. Eyke Hüllermeie, Pr. Eric Feron, professor Henry Prade, Pr. Gille Richard and Dr. Sébastien Desterke for serving as my committee members even at hardship. I also want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

I would especially like to thank Richard Alligier for his valuable advices, criticisms and four years of enjoyable collaboration. I particularly want to thank David Gianazza and Pascal Lezaud with special thanks to Hamidreza Khaksa, Masoud Ebadi Kivaj, Charlie Vanaret without forgetting all other members of the MAIAA Laboratory.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all of the sacrifices that you have made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank Sadat Mirkamali, Sadegh Razmohhesseini and all of my friends who supported me in the organizations of my Phd dissertation. Last but not the least, I would like express appreciation to my beloved fiancé without whose love, encouragement and assistance, I would not have finished this thesis.

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

To respond to the increasing levels of air traffic demand, we need Air Traffic Management (ATM) systems capable of automatically detecting and solving potential aircraft trajectory conflicts. The efficiency of these conflict solvers are mainly influenced by their trajectory prediction module. Even if most aircraft have very efficient control modules which are able to determine precisely their positions in the future, this information is not available to ground control systems. Ground control systems have access to the past positions of the aircraft and some forecast information. Moreover, the lack of critical information such as the mass of the aircraft makes the use of physical models very tricky. Thanks to the monitoring and storage of ground control data on large period, the use of statistical regression methods to predict the future positions of the aircraft trajectory appears to be a reliable solution. However, it is unrealistic to expect that statistical techniques will provide precise prediction due to the lack of some important information (such as ground control orders which are not recorded by computers). Because a safe and efficient aircraft trajectory predictor is a prerequisite for the implementation of automated tools that detect and solve trajectory conflicts, it seems more reasonable to predict intervals rather than precise aircraft positions. The thesis considers multiple topics. First, we focus on the tools for representing the uncertainty around the prediction. Next, we provide a review of the state of the art on interval prediction methods in regression and proposes a framework for comparing and checking the reliability of these methods. Indeed, what we propose is an interval prediction method which generally provides a smaller reliable prediction envelope. We finally apply it to our aircraft trajectory prediction problem.

Context of the thesis

There are different kinds of regression techniques which estimate different characteristics of the conditional distribution of the response variable $Y(x)$. The most common approaches estimate the mean of the random variable $Y(x)$ and are usually known as least-squares techniques. Robust regression approaches are similar to least-squares techniques but they are designed to be robust to outliers and violations of the least-squares assumptions. Another kind, called quantile regression, estimates the conditional quantiles of the response variable. In each category, the regression function can be estimated with a parametric linear, a parametric non-linear or a non-parametric method. This results in linear, non-linear or non-parametric regression models. These models are always built with finite sample

sizes, thus the predicted mean or quantile is an estimate of the true unknown conditional mean or quantile of the random variable $Y(x) = f(x) + \varepsilon$. Therefore while dealing with finite size datasets, we need to make some statistical inferences. In this work we are interested in finding two-sided prediction intervals in regression models which contain, with a high confidence level, at least a desired proportion of the conditional response variable. Such interval prediction models can be obtained with tolerance intervals for regression or confidence interval on quantile regression, but the application of these methods in the non-linear and particularly the non-parametric case are limited in the literature.

We can divide interval prediction approaches into two categories: The first category methods are based on the estimated conditional mean. These methods are usually based on least-squares models and propose interval prediction techniques that are centered on the estimation of the mean regression function. These approaches generally assume a non-biased regression model with a Gaussian error having constant variance. On the other hand we have quantile regression methods which directly estimate these intervals. Quantile regression methods are more robust to outliers and have less assumptions than the least-squares approaches. But they suffer from other weaknesses like slower speed of convergence and the crossing quantile effect.

The discussed interval prediction methods are in the classical frequentist statistics framework. However the interval prediction problem is not restricted to this framework. The uncertainty concept is divided into two types: the first uncertainty is due to fluctuations or heterogeneity of materials and components space and time, because of the intrinsic stochastic variability of individuals, materials and components. This type of uncertainty is known as “aleatory uncertainty” which shows its relation to the randomness in gambling and games of chance. The second, known as “epistemic uncertainty”, arises from observation errors, censoring, hidden nature of the system, lack of variables and scientific ignorance. This type of uncertainty can usually be reduced by additional observations and further empirical effort. When the uncertainty about quantities is just aleatory, probability theory is the ideal framework. However for situations in which the uncertainty about quantities contain both the aleatory and epistemic uncertainties, different competing approaches have been proposed. One idea states that the classical probability theory can be addressed in both the uncertainty types but many authors disagree. Several works have addressed the concept of modeling both the epistemic and aleatory uncertainties by using probability theory and they resulted in similar ideas which mainly state that one can use bounds on probability instead of precise probabilities. This idea was initiated by Boole [Boole 54] and has been developed by Walley and Fine [Walley 82], Williamson [Williamson 89] and Berleant [Berleant 93]. This modeling brings several new uncertainty frameworks such as: p-box, possibility theory and Transferable Belief Models (TBM).

Propositions and Contributions

As stated before, our work is based on the classical frequentist probability framework, but we do not restrict it to aleatory models. So we also propose a possibilistic representation of our statistical models which lets us access the wide community of imprecise probabilities. The possibility theory provides the simplest uncertainty framework which can be used to represent imprecise or incomplete knowledge. A quantitative possibility distribution needs at most $n-1$ values to fully represent the possibility distribution of a sample set of n observations [Destercke 08]. Moreover, a possibility distribution is an appropriate uncertainty model for encoding two-sided statistical confidence intervals or credible intervals for future realizations from an unknown or partially known probability distribution. The possibility distribution contains all the probability distributions that are respectively upper and lower bounded by the possibility and the necessity measure [Didier 06]. Therefore the possibility choice corresponds fully to the aim of this thesis which is to provide robust two-sided intervals for future aircraft positions. One major contribution of our work addresses the high confidence two-sided interval prediction problem. For a given sample set, there are different methods for building possibility distributions which encode the family of probability distributions that may have generated our sample set. Apart from our recent study [Ghasemi Hamed 12b], all the existing methods are based on parametric and distribution free confidence bands. In this work, we look at these new possibility distributions. These distributions encode different kinds of uncertainties that have not been treated before. They encode statistical tolerance and prediction intervals. We also propose a possibility distribution encoding the confidence band of the normal distribution, which improves the existing ones for all sample sizes. These distinct possibility distributions can be used to build different types of possibilitic regression for the same sample set. These possibilitic regression models are the result of exploiting the relationship between statistical inference on regression models and the possibility theory.

Once we have chosen our uncertainty framework and studied different types of confidence intervals, we can focus on the high confidence interval prediction problem in regression models. We refer to such methods as "interval prediction methods". One of our contributions is the review and the comparison of different least-squares and quantile regression techniques used to find intervals which contain a desired proportion of the response variable. We take advantage of this work to address common mis-understood questions about interval prediction methods in the machine learning community. We explain their applications and review their drawbacks. As pointed out at the beginning paragraph, we are interested in finding intervals in regression models which contain, with a high confidence level, at least a desired proportion of the conditional response variable. For that purpose, we introduce a new type of interval prediction method named "predictive interval methods". A predictive interval model contains, for any query point x, at least a desired proportion of the conditional distribution of the response variable. Such models can be obtained with tolerance intervals for regression or confidence interval on quantile regression, but these concepts have limited applications in the literature. We propose predictive interval models for local

linear regression models. Our predictive interval models are applied for two-sided interval prediction, however one can easily extend them to a one-sided interval prediction context. Then, we introduce a statistical test to check if an "interval prediction model" is a "predictive interval model". In the same context, we introduce two measures for ranking interval prediction models. These measures rate the efficiency and the tightness of the obtained envelope.

Our main contribution is to propose two predictive interval methods for non-parametric regression. Our local linear predictive interval methods are based on the local linear regression and give variable size intervals. We assume that the mean regression function is locally linear and that the prediction error is locally homoscedastic (heterocedastic in general). Our method does not neglect the regression bias and finds intervals that work properly with biased regression models. The proposed predictive intervals are based on the leave-one-out or 10-fold cross-validation prediction errors of the local linear regression. We also briefly discuss the concept of simultaneous predictive intervals. A simultaneous predictive interval model provides simultaneous predictive intervals for all the points in the predictor space, $\forall x \in \mathcal{X}$. β-content simultaneous predictive intervals can be obtained with simultaneous tolerance intervals for regression in linear regression. This work introduces simultaneous predictive intervals for K-Nearest Neighbor (KNN) regression. It is similar to predictive intervals with local linear regression but has three main differences: first, it is performed in a simultaneous context. Second, it uses a KNN regression method instead of a local linear. Finally, the simultaneous predictive interval for the response value is obtained directly with the observation values instead of prediction errors.

In order to validate our findings, we use several regression datasets to compare our predictive interval method for local linear regression with other interval prediction methods. The selected methods will be tested on their capacity to provide two-sided β-content predictive interval models. The models are compared by their reliability, efficiency, precision and the tightness of their obtained envelope. This comparison is made regardless to any variable selection or outliers detection preprocessing. We also take advantage of our evaluation chapter to show that the conventional interval prediction method is not appropriate for high confidence interval prediction. It is almost always less efficient than our predictive interval methods and their envelope is almost always larger than the envelope obtained by our methods. After a detailed comparison of the methods, we come back to the ground-based aircraft trajectory prediction which was our first motivation. In this work, a standard point-mass model and statistical regression method is used to predict the altitude of climbing aircraft. In addition to the standard linear regression model, we use two common non-linear regression methods, Least Squares Support Vector Machines (LS-SVM) and the Loess method. These methods lead to five different prediction models and they are compared, based on their point based prediction performance. However because of the critical nature of our problem and regarding the safety constraints, it seems more reasonable to predict intervals rather than precise aircraft positions. We apply nine different interval prediction methods to our aircraft trajectory prediction dataset. Some of these interval prediction models are built on the obtained prediction models and others (quantile regression based models)

are constructed without using the preceding regression models. Our experiments compare these models based on their reliability, efficiency and tightness of the obtained envelope.

Organization

This work is divided into three parts and ten chapters. The first part contains the first three chapters. It reviews imprecise probabilities and discusses the problem of interval prediction within the statistics and possibility theory. The second part is dedicated to the interval prediction problem within the statistical regression context. This part briefly discusses the relationship between all the mentioned statistical interval prediction methods and possibilistic regression with crisp input and output data. This part is composed of Chapters 4 to 8. The third part contains Chapters 9 and 10 which describe our experiments.

Chapter 1 gives a brief review of the mostly used uncertainty frameworks that address both aleatory and epistemic uncertainty and have been used within the regression context.

Chapter 2 reviews some of the most classical confidence sets in frequentist statistics. Tolerance intervals are explained in 2.3 and they are the core concept of our work. A tolerance interval depends on the number of observations that was used to construct it. Thus it is not an asymptotic interval and this is what makes it an interesting tool for statistical inference based on finite sample size. Chapter 3 uses the possibility framework to encode a family of probability distributions which may have generated our sample set. We have partially published this chapter in [Ghasemi Hamed 12b] and [Bounhas 13].

Chapter 4 is a background of the regression analysis with an exhaustive state of the art on fuzzy and non-fuzzy interval prediction methods.

Chapter 5 covers interval prediction with statistical regression. The contribution of this chapter is to review and compare different least-squares and quantile regression techniques used to find such intervals. We address a mis-understood interval prediction method in the machine learning community. We explain its applications and review its drawbacks.

Chapter 6 introduces a new interval prediction framework within the regression context. This chapter introduces the concept of regression predictive intervals and regression predictive interval models. This concept is followed by a test to verify if an "interval prediction model" is a "predictive interval model". We also describe the relationship between predictive intervals models and tolerance intervals for regression and confidence interval on quantile regression. We explain how to choose a confidence level γ to obtain efficient and reliable predictive interval models. The final part is dedicated to an illustrative example which compares two distinct interval prediction methods on the motorcycle dataset [Silverman 85].

Chapter 7 deals with predictive interval methods for local linear regression. This chapter begins by proposing a method to compute tolerance intervals for local linear regression. We describe how to use the tolerance intervals to obtain predictive interval models and then, show how to obtain our interval prediction models with a commonly used local linear regression method called loess. This chapter ends up with an illustration section and a conclusion part which compares existing methods.

Chapter 8 introduces the concept of simultaneous predictive intervals. We introduce simultaneous predictive interval for KNN regression. This chapter discusses briefly these intervals, but the interested reader can find more details in [Ghasemi Hamed 12c]. The reader can also find a related study under the possibility theory [Ghasemi Hamed 12a].

Chapter 9 evaluates the performance of our predictive interval method for local linear regression. The selected methods are tested on their capacity to provide two-sided β-content predictive interval models. This chapter is organized in five sections: the first section describes our datasets, the second section describes the interval prediction methods used in the third section. The fourth section explains our experiments on the simultaneous predictive models which are also published in [Ghasemi Hamed 12c].

Chapter 10 is a ground-based aircraft trajectory prediction example which has been partially published in [Ghasemi Hamed 13]. As stated before, the main goal of this thesis is to obtain interval prediction models able to provide intervals that, with a high confidence level, contain at least a desired proportion of the distribution for the future aircraft position. The experiments part compares our predictive interval for Loess with other point mass based and regression based interval prediction models.

Part I

Uncertainty Modeling

Chapter 1

Uncertainty Frameworks

Contents

Risk analysis contains two different types of uncertainties [Ferson 03]: the first one is an uncertainty which is due to fluctuations or heterogeneity of materials and components space and time, because of the intrinsic stochastic variability of individuals, materials and components. This type of uncertainty is known as "aleatory uncertainty" which shows its relation to the randomness in gambling and games of chance. It is also called as "irreducible uncertainty" because, by definition, one cannot reduce the aleatory uncertainty by additional empirical study. The second known as "epistemic uncertainty" arises from observation error, censoring, hidden nature of the system, lack of variables and scientific ignorance. This type of uncertainty can usually be reduced by additional observations and further empirical effort. When the uncertainty about quantities is just aleatory, probability theory is the ideal framework. However for situations in which the uncertainty about quantities

contains both the aleatory and epistemic uncertainties, different competing approaches have been proposed. The first approach states that the classical probability theory can be addressed in both the uncertainty types. Shafer [Shafer 76] argued that an approach that takes into account the indistinguishability of underlying states within bodies of evidence would be required. Walley [Walley 91], proposed that this problem must be treated by the imprecise probabilities theory and Williamson [Williamson 89] and Williamson and Downs [Williamson 90] investigated arithmetics on p-boxes. Smets introduced the Transferable Belief Model [Smets 94] as an interpretation of the Dempster-Shafer model [Shafer 76]. In the same context, Dubois [Didier 06] proposed a possibility distribution as a family of probability distributions. Destercke et al. [Destercke 08] introduced a generalized form of p-boxes which have interesting connections with other well known uncertainty representations. Apart from the Walley [Walley 91] book, there is little rigorous and detailed work that compares these uncertainty frameworks in a concise manner. Meanwhile, the Destercke et al. [Destercke 08] study gives a brief introduction and review of these subjects.

The current thesis is based on the classical frequentist probability and the possibility theory [Zadeh 78], but we also take a brief and non-exhaustive review of the mostly used uncertainty frameworks that address both aleatory and epistemic uncertainty and have been used within the context of regression. In Chapter 3 we use the possibility framework to encode a family of probability distributions which may have generated our dataset.

1.1 Imprecise probabilities

Walley's book [Walley 91] is a reference work for the theory of imprecise probabilities. He modeled uncertainty by lower and upper bounds (called coherent lower previsions) on the expected value of bounded real-valued functions on the random variable X. Imprecise probability theory is a very general concept. From a mathematical point of view it involves all the uncertainty models represented in this work. Imprecise probability gives appropriate and encompassing ways to treat several of the most practical uncertainty models and risk analysis problems as described by [Ferson 03]:

- Partially or imprecisely specified distributions;
- Inconsistency in the input data quality;
- Model uncertainties;
- Lack of sufficient knowledge on dependencies;
- Non-stationarity in distributions;
- Consequential measurement uncertainties;
- Small sample sizes.

Several works have addressed the concept of modeling both the epistemic and aleatory uncertainties by using probability theory and they resulted in similar ideas which mainly state that one can use bounds on probability instead of precise probabilities. These idea was first initiated by Boole [Boole 54] and has been developed by Walley and Fine [Walley 82], Williamson [Williamson 89] and Berleant [Berleant 93]. Note that Walley's definition of imprecise probabilities share some similarities with the classical robust statistics, but it is not based on the same principles described in [Huber 09].

1.1.1 Bayesian Interpretation

Bayesian statistical inference models beliefs and preferences with precise probability distributions (priors), and then it makes use of the Bayes rule to combine these priors with statistical data. We distinguish objective Bayes theory from subjective Bayes theory. The objective Bayes theory began with Bayes work [Bayes 63] and was developed by Laplace [Laplace 12] [Laplace 14]. The idea behind these theories is to use "non-informative" priors to model the ignorance of any prior probability distribution. Criticism of the objective Bayesian theory is detailed Chapter 2 in [Fisher 59], Chapter 4 in [Savage 72] and sections 5.1.2 , 5.5 and 7.4 in [Walley 91]. Subjective Bayesian theory is a more popular version of the Bayesian approach. It suggests that probability distribution models the personal belief [De Finetti 72].

Bayesian sensitivity analysis uses some kind of inference which is the same as the imprecise probability theory. In Bayesian sensitivity analysis [Berger 84], the analyst makes several precise Bayesian inferences with different precise probability priors. This produces a range of precise posterior probability distribution which leads to a range of probability measures or a range of expected utilities in decision making. More details and discussions can be found in [Good 62, Good 65], [Huber 09] and [Walley 91].

1.1.2 Frequentist Interpretation

Huber and Strassen [Huber 73] and Huber [Huber 09] studied a frequentist interpretation of lower and upper probabilities in robust statistics. Wally and Fine [Walley 82] presented a frequentist theory of statistics to introduce upper and lower probabilities (interval-valued probability). They consider models based on independent and identically distributed observations (IID) for unlinked repetitions of experiments which are described by Interval-Valued Probability (IVP). They also suggest several generalizations of standard concepts of independence, asymptotic certainty and estimability. The idea is that we dispose of a lot of geological, economic, medical, psychological and sociological observations time series data for which we have little information concerning dependence between their observations. Such problems are modeled by non-stationary probability models, which are usually complex and are often not based on our understanding of the phenomenon. However these behaviors can be much more simply modeled by non-additive IID models and are precise enough to give useful predictions. Another type of upper and lower probability which models stationary processes having unstable time average can be found in [Grize 87] and [Fine 88].

1.2 P-box

Let P be a probability measure on the random variable X on $\mathbb{R}$ and $F(\cdot)$ be its cumulative distribution $F(x) = P(X \leq x)$. Let $\overline{F}(\cdot)$ and $\underline{F}(\cdot)$ denote two cumulative distributions for X such that for all $x, \underline{F}(x) \leq F(x) \leq \overline{F}(x)$. Then the pair $[\underline{F}(\cdot), \overline{F}(\cdot)]$ is a "p-box" for X [Ferson 03]. It means that the cumulative distribution $F(\cdot)$ is unknown but we know that it is contained in the p-box $[\underline{F}(\cdot), \overline{F}(\cdot)]$. Therefore $\underline{F}(x)$ is a lower bound on $F(x)$. It can be calculated from a lower probability measure $\underline{P}$ for the random variable X [Walley 91]:

$$\underline{F}(x) = \underline{P}(X \leq x)$$

and the upper bound can be obtained by

$$\overline{F}(x) = 1 - \underline{P}(X > x).$$

Probability box (or p-box) is a a framework for modeling both the aleatoric and epistemic uncertainties. This is often used in risk analysis or uncertainty modeling where numerical calculations must be performed. Probabilistic knowledge from experts is usually represented by cumulative distributions [Technology 91], and so the p-box can benefit from such tools but it also offers the opportunity to have epistemic uncertainties.

Williamson [Williamson 89] and Williamson and Downs [Williamson 90] investigated arithmetics on p-boxes. They described detailed examination of numerical methods for calculating the distribution of arithmetic operations on pairs of p-boxes. Note that there is no general relationship between the frameworks of possibility distributions, p-boxes and probability intervals. Comparison of possibility distributions and p-boxes can be found in [Baudrit 06]. Destercke et al. [Destercke 08] defined a generalized form of p-boxes which have interesting connections with other well known uncertainty representations. They show that generalized p-boxes are equivalent to pairs of possibility distributions, and that they are special kinds of random sets. They also present a review of common uncertainty representation frameworks, their relationships and their transformations.

In the same context, we have probability intervals [de Campos 94] which are lower and upper bounds of probability distributions. They are defined by a set of intervals $L = \{[l(x), u(x)], \forall x \in \mathcal{X}\}$ where

$$l(x) \leq p(x) \leq u(x), \forall x \in \mathcal{X}, \text{ and } p(x) = P_X(x).$$

and $\mathcal{X}$ is the domain of x. There is a particular case of lower and upper probabilities where the constraints can only affect individual probabilities x. Restriction which affect more than one individual probability like $P(x_1) + P(x_2) \leq u_{ij}$ are possible in lower and upper probabilities, but they are not permitted in probability intervals [de Campos 94]. Probability intervals are suitable for modeling uncertainties on multinomial data where they can be used to represent lower and upper confidence bounds. [Destercke 08]

1.3 Possibility Theory

In 1978, Zadeh introduced the possibility theory [Zadeh 78] as an extension of his theory of fuzzy sets. Possibility theory offers an alternative to the probability theory when dealing with some kinds of uncertainty. The possibility theory has a qualitative and a quantitative interpretation. Despite both interpretations sharing the same elementary notions, qualitative and quantitative possibility theories diverge on conditioning and combination tools. Qualitative possibility theory has a close link to non-monotonic reasoning whereas quantitative possibility involves notions similar to the probability theory. Quantitative possibility is an imprecise probability framework that represents probability bounds and it can also be seen as a special case of belief functions [Dubois 98]. Quantitative possibility distributions can also be viewed as a family of probability distributions. Then, the possibility distribution contains all the probability distributions that are respectively upper and lower bounded by the possibility and the necessity measure [Didier 06]. For a given sample set, there are different methods for building possibility distributions which encode the family of probability distributions that may have generated the sample set [Ghasemi Hamed 12b, Aregui 07b, Masson 06, Aregui 07a]. The mentioned methods are almost all based on parametric and distribution free confidence bands.

1.3.1 Definition

Possibility theory [Zadeh 78, Dubois 80], was initially created in order to deal with imprecision and uncertainty due to incomplete information. This kind of uncertainty may not be handled by probability theory, especially when a priori knowledge about the nature of the probability distribution is lacking. In possibility theory, we use a membership function π to associate a distribution on sample space Ω. In this paper, we only consider the case $\Omega = \mathbb{R}$.

Definition 1 *A possibility distribution π is a function from Ω to $[0,1]$ ($\pi : \mathbb{R} \to [0,1]$).*

The definition of the possibility measure Π is based on the possibility distribution π such that:

$$\Pi(A) = \sup(\pi(x), \forall x \in A). \tag{1.1}$$

The necessity measure is defined by the possibility measure

$$\forall A \subseteq \Omega, N(A) = 1 - \Pi(A^C) \tag{1.2}$$

where A^C is the complement of the set A. A distribution is normalized if: $\exists x \in \Omega$ such that $\pi(x) = 1$. When the distribution π is normalized, we have:

$$\Pi(\emptyset) = 0, \Pi(\Omega) = 1,$$
$$\Pi(U \cup V) = \max(\Pi(U), \Pi(V))$$
$$N(A) \leq \Pi(A).$$

Based on Zadeh's [Zadeh 78] consistency principle of possibility "what is probable should be possible", and by considering the definition of necessity, we obtain the following inequalities:

$$N(A) \leq P(A) \leq \Pi(A), A \subset \Omega. \tag{1.3}$$

Thus by using the possibility and necessity measures, like in the Dempster-Shafer theory, we can define upper and lower values to describe how likely an event is to occur. Note that for any event A, we will have either :

$$\Pi(A) = 1 \text{ or } N(A) = 0$$

which means that the pair $[N(A), \Pi(A)]$ will be $[0, \alpha]$ or $[\beta, 1]$.

Definition 2 *The α-cut A_α of a possibility distribution $\pi(\cdot)$ is the interval for which all the points x located inside it have a possibility membership $\pi(x)$ greater than or equal to α.*

$$A_\alpha = \{x | \pi(x) \geq \alpha, x \in \Omega\}, \tag{1.4}$$

Therefore, we have:

$$[N(A_\alpha), \Pi(A_\alpha)] = [\alpha, 1].$$

1.3.2 Probability-possibility transformation

In many cases it is desirable to move from the probability framework to the possibility framework. This is why several transformations based on various principles such as consistency (this principle states that "what is probable should be possible") or information invariance have already been proposed [Civanlar 86, Delgado 87, Klir 90, Dubois 93a, Dubois 04]. Dubois et al. [Dubois 93b] suggest that when moving from the probability to possibility framework, we should use the "maximum specificity" principle which aims to find the most informative possibility distribution. The definition of the maximum specificity principle is based on three other concepts known as "inter-quantile", "smallest β-content interval" and "specificity". W begin by their definition and then formally define the maximum specificity principle. We denote the density function of a probability distribution by $f(\cdot)$, its receptive cumulative distribution function (cdf) by $F(\cdot)$ and its probability measure by P.

Definition 3 *The interval between the lower and upper quantiles of the same level are called inter-quantiles. The inter-quantile at level p is defined by*

$$[F^{-1}(p), F^{-1}(1-p)], 0 < p < 0.5 \tag{1.5}$$

where $F^{-1}(\cdot)$ is the inverse function of the continuous strictly-monotone cdf $F(\cdot)$.

An inter-quantile at level p contains β proportion of the distribution where $\beta = 1 - 2p$. We will call a β-content inter-quantile I_β, the interval that contains β proportion of the underlying distribution, we have $Pr(I_\beta) = \beta$.

Definition 4 *Given a probability density function $f(\cdot)$ with a finite number of modes, we define the interval I^*_β defined below as the "smallest β-content interval" of $f(\cdot)$.*

$$I^*_\beta = \{x : f(x) \geq c\} \tag{1.6}$$

where c is defined as $\int \mathbb{1}_{I^*_\beta}(x) f(x) dx = \beta$ and $\mathbb{1}_{I^*_\beta}(x)$ is the indicator function for I^*_β.

Thus we have:

$$P(X \in I^*_\beta) = \beta$$

and I^*_β does not exists for f having infinite number of modes, for instance the uniform distribution.

Definition 5 *A possibility distribution π_1 is more specific than the possibility distribution π_2 if and only if:*

$$\forall x \in \mathcal{X},\ \pi_1(x) \leq \pi_2(2), \tag{1.7}$$

where $\mathcal{X}$ is the domain of x. This can also be represented by $\pi_1 \leq \pi_2$.

Definition 6 *Given the maximum specific possibility distribution (m.s.p.d) π^* that encodes the probability distribution function F (i.e. $\forall A \subseteq \Omega, N^*(A) \leq P(A) \leq \Pi^*(A)$) we have, for all π such that $\forall A \subseteq \Omega, N(A) \leq P(A) \leq \Pi(A)$, $\pi^*(x) \leq \pi(x), \forall x \in \Omega$.*

Because the possibility distribution explicitly handles the imprecision and is also based on an ordinal structure rather than an additive one, it has a weaker representation than the probability one. This kind of transformation (probability to possibility) may be desirable when we are in presence of weak source of knowledge or when it is computationally harder to work with the probability measure than with the possibility measure. The "most specific" possibility distribution is defined for a probability distribution having a finite number of modes [Dubois 04]:

$$\pi^*(x) = sup(1 - P(X \in I^*_\beta), x \in I^*_\beta) \tag{1.8}$$

where π_t is the most specific possibility distribution, I^*_β is the smallest β-content interval [Dubois 04]. Then, in the spirit of equation 1.10, given f and its transformation π^* we have:

$$A^*_\alpha = I^*_\beta \quad \text{where } \alpha = 1 - \beta.$$

Figure (1.1) presents the maximum specific transformation (in blue) of a normal probability distribution (in green) with mean and variance respectively equal to 0 and 1 ($\mathcal{N}(0,1)$). It also illustrates two inter-quantiles of the standard normal distribution and their respective α-cuts in the m.s.p.d for $\mathcal{N}(0,1)$.

Lemma 1 *The maximum specific possibility distribution (m.s.p.d) $\pi^*(\cdot)$ of the unimodal symmetric probability density function $f(\cdot)$ can be built as follows:*

$$\pi^*(x) = \begin{cases} 2F(x) = 1 & \text{if } x = \mu, \\ 2F(x) & \text{if } x < \mu, \\ 2F(2\mu - x) & \text{if } x > \mu, \end{cases}$$

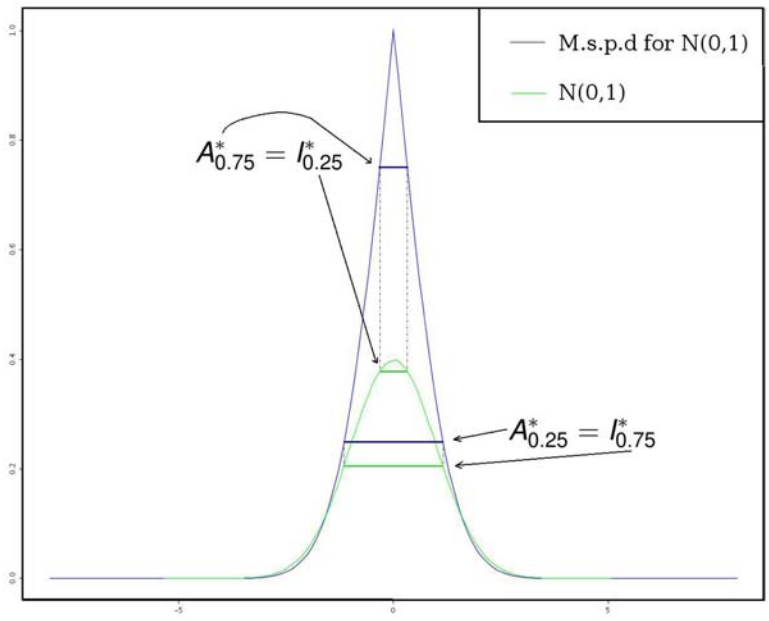

Figure 1.1: Illustrating 0.25-cut and 0.75-cut and their respective smallest β-content intervals.

Where, μ denotes $f(\cdot)$'s mode or median and $F(\cdot)$ its cumulative distribution function.

Proof : The mode, median and mean of a symmetric and unimodal distribution are identical, so *the smallest β-content interval I^*_β of $f(\cdot)$ is also its inter-quantile at level* $\frac{1-\beta}{2}$. Thus the smallest inter-quantile I^*_β containing x is the inter-quantile that maximizes $1 - P(X \in I^*_\beta)$ in Equation 1.8 with x being on its borders. If x is smaller than the mode($x < \mu$), then x is equal to the I^*_β's lower bound $x = F^{-1}(\frac{1-\beta}{2})$. Otherwise it is equal to the I^*_β's upper bound $x = F^{-1}(1 - \frac{1-\beta}{2})$. By simple calculations and application of Equation 1.8 we deduce the previous proposition. ■

Proposition 1 *The maximum specific possibility distribution (m.s.p.d) $\pi^*(\cdot)$ of a unimodal symmetric probability density function $f(\cdot)$ can be built by calculating the β-content inter-quantile I_β of $f(\cdot)$ for all the values of β, where $\beta \in [0, 1]$.*

Proposition (1) is a direct result of Lemma (1).

1.3.3 Encoding a family of probability distributions

One interpretation of possibility theory, based on Zadeh's [Zadeh 78] consistency principle of possibility ("what is probable should be possible"), is to consider a possibility distribution as a family of probability distributions (see [Didier 06] for an overview). Thus, a possibility

distribution π will represent the family of the probability distributions Θ for which the measure of each subset of Ω will be bounded by its possibility measures:

Definition 7 *A possibility measure* Π *is equivalent to the family* Θ *of probability distributions* F *such that*

$$\Theta = \{F | \forall A \in \Omega, P(A) \leq \Pi(A)\}, A \subseteq \Omega. \tag{1.9}$$

Now let θ be a set of cdfs, where F is defined by a possibility distribution function $\pi(\cdot)$. Then, an alternative to equations (1.9) is:

$$\forall \alpha \in [0,1], \forall F \in \Theta, I^*_{F,\beta} \subseteq A_{\pi,\alpha}, \tag{1.10}$$

where $\beta = 1 - \alpha$ and $A_{\pi,\alpha}$ is the α-cut of possibility distribution $\pi(\cdot)$. Thus, a possibility distribution encodes a family of probability distributions for which each smallest $(1-\alpha)$-inter-quantile is bounded by a possibility α-cut. This is stated formally as:

Proposition 2 *Let* Θ *be a family of probability and let* $\pi(\cdot)$ *denote the possibility distribution function encoding* Θ*. Each* α*-cut of* $\pi(\cdot)$ *contains the smallest* $(1-\alpha)$*-content inter-quantile of all of the probability distributions included in family* Θ*.*

This property makes possibility distributions a good tool for representing two-sided confidence intervals on nested random sets. *In other words, a possibility distribution is an appropriate uncertainty model for encoding two-sided statistical confidence intervals or credible intervals for future realizations from an unknown or partially known probability distribution.*

1.4 Transferable Belief Model (TBM)

The Transferable Belief Model (TBM) [Smets 94, Smets 13] is an interpretation of the Dempster-Shafer model [Shafer 76], and it is used for representing quantified beliefs with belief functions. TBM is a model for point-wise quantified beliefs. TBM has two levels: the first level is known as the "credal" level in which one uses belief functions to represent its belief on world. The second level is the "pignistic" level where the decision making takes place. Whenever we need to make a decision, the belief functions are transformed using "pignistic transformation" to probability functions. TBM is widely used as a formal framework for information fusion [Aregui 07c, Quost 07, Mercier 08] and is used increasingly in imprecise data analysis [Masson 04, Denoeux 04, Petit-Renaud 04, Su 13]. Some important aspects of the Transferable Belief Model are the following [Smets 94]:

- This is a two level model; beliefs are represented and updated in the credal level and a pignistic level is used only to make decisions.
- TBM departs from the idea that we do not require any probability function even though it may exist. There is no need to have a probability distribution to obtain a

belief function. The TBM is suited for subjective models and personal beliefs. This is the application domain of the Bayesian framework. The fundamental contrast between the Transferable Belief Model and the Bayesian framework is its complete dissociation from any probability function.

- The credal level always precedes the pignistic level.

Let $\mathbf{X} = \{x_1, ..., x_n\}$ be a finite set and X be a variable with Ω as its domain. TBM encodes the knowledge of variable X by a so-called basic belief assignment (bba) m which is a mapping from 2^Ω to $[0, 1]$ such that:

$$\sum_{A \subseteq 2^\Omega} m(A) = 1.$$

Each mass $m(A)$ shows our belief to the statement that variable X can have a value inside A, $X \subset A$. The difference between basic belief assignment (bba) and probability models is that masses can be given to any subset of Ω instead of just assigning mass to atoms of Ω. Subsets of Ω that have a mass greater than zero $m(A) > 0$ are called the focal sets of m. When these focal sets are nested, the bba m is known to be consonant. The belief $bel(\cdot)$, plausibility $pl(\cdot)$ and commonality $q(\cdot)$ functions are defined as below:

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \tag{1.11}$$

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{1.12}$$

$$q(A) = \sum_{B \supseteq A} m(B), \tag{1.13}$$

and these formulas hold for all $A \subseteq \Omega$.

When m is consonant, the plausibility function is also a possibility measure and the corresponding possibility distribution denoted by $\pi(\cdot)$ is defined by:

$$\forall x \in \Omega,\ \pi(x) = pl(x) = q(x)$$

where the possibility and commonality function respectively verify the following two properties:

$$\forall A, B \subseteq \Omega,\ pl(A \cup B) = max(pl(A), pl(B)),$$
$$\forall A, B \subseteq \Omega,\ q(A \cup B) = min(q(A), q(B)).$$

The necessity measure and possibility measure are, respectively, particular cases of belief functions and plausibility functions where they are induced from a random set with nested focal sets. Having a possibility distribution $\pi(\cdot)$, one can obtain its corresponding bba m in the TBM model as follows [Aregui 07a]:

Let $\pi_n = \pi(x_n)$, and consider an arrangement of the elements of Ω such that:

$$\pi_1 \geq \pi_2 \geq \cdots \geq \pi_n.$$

Then, the corresponding bba is defined as follows:

$$m(A) = \begin{cases} 1 - \pi_1, & \text{if } A = \emptyset, \\ \pi_k - \pi_{k+1}, & \text{if } A = \{x_1, \cdots, x_n\}, \text{ where } k \in \{1, \cdots, n\}, \\ \pi_n, & \text{if } A = \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{1.14}$$

For a more detailed comparison of TBM and the possibility theory see [Smets 90a].

Having a sample of observations, if we assume that our data comes from a random variable X having an unknown probability distribution P_X, we would like to express our beliefs about future realizations of X from the sample set. The inferred belief function must be such that its pignistic probability distribution is equal to P_X. Masson et al. [Masson 06], suggested simultaneous confidence intervals of the multinomial distribution to build possibility distributions and Denoeux [Denoeux 06] applied the same concept in the TBM framework. Aregui et al. [Aregui 07a] proposed a method for building a TBM belief function from a random sample drawn from a Gaussian distribution. In the same context, Aregui et al. [Aregui 07b] used the Kolmogorov confidence band [Birnbaum 52] to construct predictive belief functions for sample sets drawn from an unknown distribution.

1.5 Confidence Intervals

1.5.1 Frequentist Confidence Interval

In frequentist statistics, a confidence interval is an estimated interval based on past observations, which states how frequently it contains the true parameters. The frequency of the confidence interval is known as its confidence level. More specifically, the term "confidence level" means that, if one uses a sample to find a confidence interval with a desired confidence level, then by repeating the same experiment, the proportion of confidence intervals constructed based on different samples (with the same number of observations and from the same distribution), which contains the true parameter will converge to the previously selected confidence level.

Definition 8 *Let $\mathbb{X} = \{X_1, \cdots X_n\}$ be a random sample from a probability distribution with unknown parameters. A confidence interval for the parameter θ, with confidence level or confidence coefficient $1 - \alpha$, is an interval determined by the pair of statistics $u(\mathbb{X})$ and $v(\mathbb{X})$, with the following property:*

$$P_\theta\Big(u(\mathbb{X}) \leq \theta \leq v(\mathbb{X})\Big) = 1 - \alpha$$

The generalization to multivariate interval estimation is the confidence region.

1.5.2 Bayesian Credible Interval

A credible interval plays the same role in Bayesian statistics that the confidence interval plays in frequentist statistics. A credible interval is an interval from the posterior probability distribution and it is used for interval estimation.

Definition 9 *Let $\mathbb{X} = \{X_1, \cdots X_n\}$ be a random sample from a probability distribution with unknown parameter θ, and let $P_\theta(\cdot|\mathbb{X})$ be the posterior distribution for the parameter θ, a confidence level $1-\alpha$-credible interval for θ from $P_\theta(\cdot|\mathbb{X})$, is an interval determined by the pair of statistics $u(\mathbb{X})$ and $v(\mathbb{X})$ where:*

$$P_\theta\Big(u(\mathbb{X}) \leq \theta \leq v(\mathbb{X})|\mathbb{X}\Big) = 1 - \alpha.$$

The generalization to multivariate interval estimation is the credible region. Credible intervals are different from confidence intervals in different ways:

- A credible interval has a more appealing interpretation than a frequentist confidence interval. A credible interval is a random variable that has a probability of $(1-\alpha)$ to contain the true parameter which itself is a random variable distributed according to the prior distribution. However, in the frequentist view, the confidence interval is a random variable and the true parameter is a fixed value. The frequentist confidence interval will include the true parameter or not, and so it contains the true parameter with probability 0 or 1. The confidence intervals probability is the limit of the fraction of confidence intervals constructed based on different samples that contain correctly the fixed unknown parameter.

- A credible interval encodes the information from the prior distribution into the estimate, while confidence intervals are based on the random samples.

However, as mentioned in 1.1.1, there is a classical debate about the consistency of objective Bayesian inference that selects a conjugate prior while we ignore any information on θ's distribution.

1.6 Conclusion

The TBM framework is the most general framework discussed here. P-boxes, possibility theory and lower and upper probabilities can all be modeled by belief functions. However

depending on the application, one framework may be more appropriate than the other. P-boxes are often used in risk analysis or uncertainty modeling where numerical calculations must be performed. They are suitable tools for representing knowledge from experts with a pair of cumulative distributions [Destercke 08].
The necessity measure in the possibility theory can be viewed as a coherent lower probability, thus its possibility distribution induces a family of probability distribution as defined in Equation (1.9). We have also seen that the necessity measure is a particular case of belief function. It is induced from a random set with nested focal sets. This means that, in several cases, possibility distributions cannot reflect all the available information. In order to fully represent the possibility distribution of a sample set, we need at most $|X| - 1$ values. This is the simplest uncertainty framework which can be used to represent imprecise or incomplete knowledge [Destercke 08]. Despite this lack of general expressive power, possibility distributions are yet suited for several applications. Beyond what has been stated, a psychological study shows that sometimes people treat the uncertainty like possibility rules [Raufaste 03]. We have also seen that the possibility distribution is an appropriate uncertainty model for encoding two-sided statistical confidence intervals or credible intervals for future realizations from an unknown or partially known probability distribution. The aim of this thesis is to provide robust two-sided intervals for future aircraft positions. With the preceding statement, possibility distributions are an appropriate choice for our problem.

Chapter 2

Statistical Intervals

Contents

In this chapter we will review some of the most classical confidence sets in frequentist statistics. First we will have a brief review of confidence bands. Then we will see how they can be constructed based on the confidence region of parameters. Tolerance intervals are explained in 2.3, and are the core concept of our work. A tolerance interval depends on the number of observations that was used to construct it. Thus it is not an asymptotic interval, and this is what makes them an interesting tool for statistical inference based on finite sample size. In Section 5.2.3 we continue the discussion of tolerance intervals within the regression context. Prediction intervals are the final type of intervals

discussed in this chapter. They are closely related to tolerance intervals and Section 2.4.2 discuss this relationship. The contribution of this chapter remains the comparison of well known and recent confidence bands and confidence regions of the normal distribution.

In the next chapter, we use a possibility distribution to encode the intervals explained here. Chapter 4 deals with inference on regression which describes similar statistical intervals to those described here, but in a regression context. Intervals discussed in this chapter and in Chapter 4 have their own specific definition and application. The goal of this chapter is to explain their statistical interpretation and their differences, and we think that this will help us to use them appropriately in uncertainty modeling.

2.1 One-sided and Two-sided Confidence Intervals

Definition 10 *Let $\mathbb{X} = \{X_1, \cdots X_n\}$ be a random sample from a probability distribution with unknown parameters. A one-sided upper confidence interval $IU_{1-\alpha}$ for the parameter θ, with confidence level or confidence coefficient $1-\alpha$, is an interval determined by the statistic $v(\mathbb{X})$, such that:*

$$IU_{1-\alpha} = (-\infty, v(\mathbb{X})], \text{ where } P_\theta\Big(\theta \leq v(\mathbb{X})\Big) \geq 1-\alpha. \tag{2.1}$$

An upper $(1-\alpha)$-confidence interval for θ will cover, *in at least* $100(1-\alpha)\%$ *of cases*, the next observation of X's distribution.

Definition 11 *Let $\mathbb{X} = \{X_1, \cdots X_n\}$ be a random sample from a probability distribution with unknown parameters. A one-sided lower confidence interval $IL_{1-\alpha}$ for the parametr θ, with confidence level or confidence coefficient $1-\alpha$, is an interval determined by the statistic $u(\mathbb{X})$, such as:*

$$IL_{1-\alpha} = [u(\mathbb{X}), +\infty), \text{ where } P_\theta\Big(\theta \geq u(\mathbb{X})\Big) \geq 1-\alpha. \tag{2.2}$$

A lower $(1-\alpha)$-confidence bound is an upper α-confidence bound. Using the statistic $u(\mathbb{X})$ defined in Equation (2.2), we obtain an upper α-confidence interval $IU_\alpha(x)$:

$$IU_\alpha = (-\infty, u(\mathbb{X})], \text{ where } P_\theta\Big(\theta \leq u(\mathbb{X})\Big) \leq \alpha.$$

The $(1-\alpha)$-lower confidence interval for θ will provide an upper limit which covers, *at most* $100\alpha\%$ *of the time*, the next observation of X's distribution. Similarly, a $(1-\alpha)$-upper confidence interval for θ will provide lower limits which cover, *at most* $100\alpha\%$ *proportion of the time*, the next observation of X's distribution. Once we know the distribution of an estimator, the procedure for obtaining one-sided or two-sided confidence intervals is almost the same. In this work, we only consider two-sided intervals but it is obvious that if one is able to construct a one-sided confidence interval for an estimator, it can also find its two-sided confidence interval.

2.2 Confidence Band

2.2.1 Definition

Definition 12 *The confidence band for a cdf $F(\cdot)$ is a function which associates to each x an interval $[L(x), U(x)]$ such that:*

$$P\Big(\forall x \in \mathcal{X}, L(x) \leq F(x) \leq U(x)\Big) \geq \gamma, \text{ where } \forall x \in \mathcal{X}, 0 \leq L(x) \leq U(x) \leq 1. \tag{2.3}$$

In frequentist statistics, a confidence band is an interval defined for each value x of the random variable X such that for a repeated sampling, the frequency of $F(x)$ located inside the interval $[L(x), U(x)]$ for all the values of X tends to the confidence coefficient γ. Note that given any γ level confidence band, we can use it to infer confidence intervals of the quantile function $Q(\beta) = F^{-1}(\beta) = \inf\{x \in R : \beta \leq F(x)\}$, for all $\beta \in (0, 1)$. In other words, the confidence band simultaneously gives confidence intervals for all $F^{-1}(\beta), \forall \beta \in (0, 1)$.

Let $I^c_{\beta_i} i \in (1, \cdots, n)$ be the γ-confidence interval of the unknown β_i-quantile $i \in (1, \cdots, n)$ of the unknown c.d.f F. The simultaneous condition is:

$$P\Big((\beta_1 \in I^c_{\beta_1}) \cap (\beta_2 \in I^c_{\beta_2}) \cap \cdots \cap (\beta_n \in I^c_{\beta_n})\Big) = \gamma. \tag{2.4}$$

Therefore such confidence intervals derived from confidence bands are *Simultaneous Confidence Intervals (SCI) for all population quantiles*. One can take advantage of this property to derive simultaneous γ-confidence intervals for β-content inter-quantiles of the unknown c.d.f $F(\cdot)$, and we will denote them by I^C_β.

2.2.2 Confidence bands based on confidence region of parameters

Suppose that K is a set of estimated c.d.f $\hat{F}$ for F, which in a repeated sampling the frequency that the function $\hat{F}$ will be equal to the true c.d.f F (if so we will have $\forall x, \hat{F}(x) = F(x)$), will tend to $1-\alpha$. This is what is in the equation (2.5). Note (2.5) and (2.3) are two different definitions which represent the confidence band concept with two different views. In the case of F being a parametric probability distribution, one can use the confidence region of its parameter vector to construct its confidence band [Cheng 83],[Frey 09]. Confidence bands built by confidence regions are described by:

$$P\Big(\exists \hat{F} \in K, F = \hat{F}\Big) = 1 - \alpha. \tag{2.5}$$

We know that if two parametric c.d.f F_{θ_1} with the parameter vector θ_1 and F_{θ_2} with the parameter vector θ_2 belonging to the same parametric family of probability distribution say $\mathcal{F}$ are equal, their parameter vectors will also be equal:

$$F_{\theta_1} = F_{\theta_2} \Leftrightarrow \forall x, F_{\theta_1}(x) = F_{\theta_2}(x) \Leftrightarrow \theta_1 = \theta_2. \tag{2.6}$$

Now suppose that K is a family of c.d.f defined over any parametric probability distribution $\mathcal{F}$ described by the vector parameter θ. Hence by using (2.6) we have:

$$\begin{aligned} \forall \hat{F} \in K, P\Big(\forall x, \hat{F}(x) = F(x)\Big) &= 1-\alpha \\ \Leftrightarrow \forall \hat{F} \in K, P\Big(\exists \hat{F} \in K, F = \hat{F}\Big) &= 1-\alpha \\ \Leftrightarrow P\Big(\exists \hat{\theta} \in R, \hat{\theta} = \theta\Big) &= 1-\alpha, \end{aligned}$$

where R is the set which contains just the parameters of family K. We can notice that R is the $1-\alpha$ confidence region of the unknown parameter vector θ of the c.d.f F which belongs to the family $\mathcal{F}$. Having a sample set coming from a parametric distribution F, a simple method for constructing a $1-\alpha$-confidence band from the $1-\alpha$-confidence region $R^{1-\alpha}$ of parameters of F is as follows :

1. Obtain a $1-\alpha$ confidence region $R^{1-\alpha}$ of parameters.
2. For each point $\theta \in R^{1-\alpha}$, and for all $x \in \mathcal{X}$ obtain the max and min values of $F_\theta(x)$ and denote them respectively by min_x and max_x.
3. The resulted band $[L(x), U(x)]$

$$[L(x), U(x)] = \{[min_x, max_x] \,|\forall x \in \mathcal{X}, min_x = \min_{\theta \in R^{1-\alpha}}(F_\theta(x)),\, U(x) = \max_{\theta \in R^{1-\alpha}}(F_\theta(x))\} \tag{2.7}$$

is a $1-\alpha$-confidence band for F.

For more details on the construction of continuous confidence band for parametric functions, the reader should refer to [Kanofsky 72] and [Cheng 83].

2.2.3 Confidence region for parameters of a normal distribution

Suppose that we have n observations $x_1, x_2, \cdots x_n$ drawn from a normal distribution with unknown mean μ and unknown variance σ^2. The $1-\alpha$ confidence region for the parameters of $\mathcal{N}(\mu, \sigma^2)$, contains a region in the two dimensional space of μ and σ^2 which has a probability of $1-\alpha$ to contain the true parameters value μ and σ^2. Arnold and Shavelle [Arnold 98] compared several methods to find such confidence regions. They used simulations with different sample sizes to compare the fraction of true parameters lying in each confidence region with its nominal value. They are all asymptotically equivalent, but for small sample sizes there are only two methods which satisfy appropriately the required confidence level. The first is a well known method that they call the Mood's exact region, defined by Equation (2.8). This method always gives a confidence region which is exactly the same as its nominal confidence level. Mood confidence regions are built by taking α_1

and α_2 such that $1-\alpha = (1-\alpha_1)(1-\alpha_2)$ where $1-\alpha$ is the confidence level of the found region and it is defined by:

$$\begin{aligned}\mathcal{R}(n,\overline{x},S) = \{(\mu,\sigma^2) : \frac{n-1}{\chi^2_{1-\frac{\alpha_2}{2},n-1}}S^2 < \sigma^2 < \frac{n-1}{\chi^2_{\frac{\alpha_2}{2},n-1}}S^2, \\ \overline{x} - \Phi_{1-\frac{\alpha_1}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + \Phi_{1-\frac{\alpha_1}{2}}\frac{\sigma}{\sqrt{n}}\}.\end{aligned} \tag{2.8}$$

Where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$ and Φ_q and $\chi^2_{q,k}$ are respectively the q^{th} quantile of the standard normal distribution and the q^{th} quantile of the chi-square distribution with k degrees of freedom. If $\alpha_1 = \alpha_2$, the confidence region is not optimal (because the chi-square distribution is asymmetric), so in the same paper [Arnold 98], they give values for the combination of α_1 and α_2 that gives the smallest possible region for a fixed confidence level $1-\alpha$ and for a fixed number of observations n. Table 2.1 is extracted from Table 4 in [Arnold 98] and gives the values for α_1, α_2 and δ. $1-\alpha_1$ denotes the confidence interval of the distribution of $\overline{x}$ which is a normal distribution, $1-\alpha_2$ is the confidence interval of the chi-square distribution and δ is the proportion of α_2 which is put in the lower tail of the chi-square distribution.

$1-\alpha$	n	α_1	α_2	σ
0.95	10	0.0117	0.0388	0.0384
0.95	25	0.0180	0.0326	0.0307
0.95	100	0.0231	0.0275	0.0219

Table 2.1: α_1, α_2 and δ values to find the smallest Mood confidence region, extracted from Table 4 in [Arnold 98].

The second method uses the likelihood ratio test to build confidence regions having a confidence level a bit smaller than the required value. The likelihood ratio confidence region has even smaller area than the Mood's optimal confidence region.

Frey [Frey 09] proposed the *minimum area confidence region* and the *minimum area confidence band* for the normal distribution. She proposed two types of confidence region, the first one is the Minimum Area (MA) confidence region and the other is the confidence region that yields the minimum area confidence band which we denote MAB confidence region. She stated that the Minimum Area (MA) confidence region is asymptotically equivalent to Cheng and Iles [Cheng 83] and other maximum likelihood confidence regions in statistics literature, however the Minimum Area confidence band improves over other confidence bands for all sample sizes. Figures (2.1 ,2.2 and 2.3) show four different confidence regions for parameters of the normal distribution. In these figures, we can see the smallest Mood confidence region, as defined in [Arnold 98], Mood confidence region [Arnold 98], Cheng and Iles [Cheng 83] and Frey's minimum area confidence region [Frey 09]. These regions are defined for sample sets with sizes $n = 10, 25$ and 100, all having $(\bar{X}, S) = (0, 1)$

and a confidence level of 0.95. It can be seen that Frey's minimum area confidence region [Frey 09] is always the smallest.

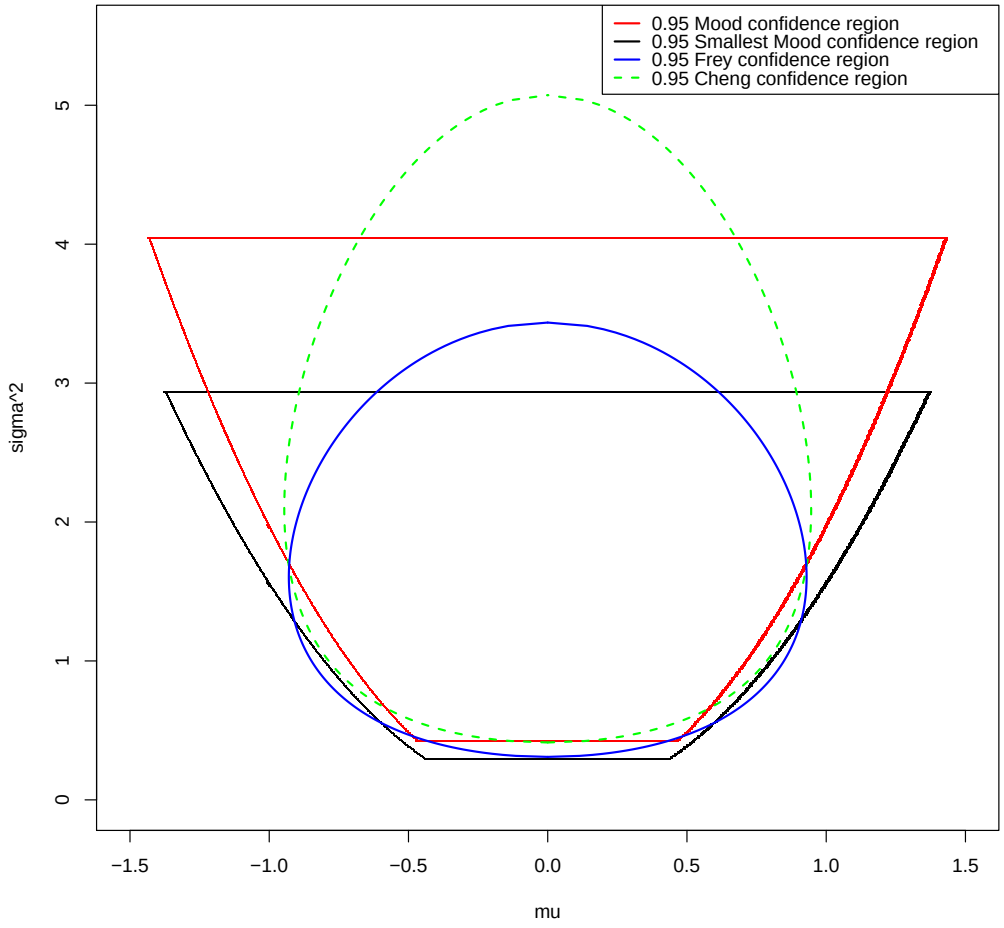

Figure 2.1: 0.95 confidence region for parameters of a normal distribution based on a sample set with size $n = 10$ and with $(\bar{X}, S) = (0, 1)$.

2.2.4 Confidence band for a normal distribution

There are already several confidence bands for the normal distribution. Some confidence bands are obtained directly [Kanofsky 72] while others are constructed based on confidence regions [Cheng 83, Frey 09]. One can also use Mood's confidence region or the Smallest Mood's confidence region along with Equation (2.7) to obtain confidence bands for the normal distribution. Frey [Frey 09] proposed its *minimum area confidence region* and *minimum area confidence band* for the normal distribution. Note that the first one is the

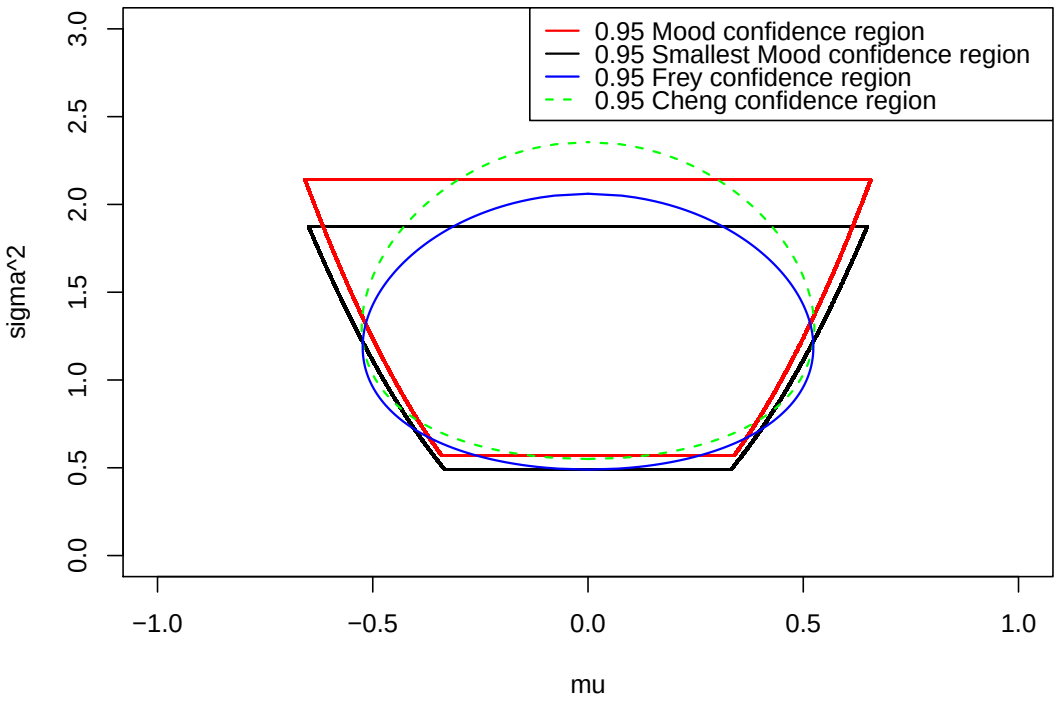

Figure 2.2: 0.95 confidence region for parameters of a normal distribution based on a sample set with size $n = 25$ and with $(\bar{X}, S) = (0, 1)$.

confidence band based on the Minimum Area confidence region denoted MAR confidence band and the second one is just the Minimum Area confidence band (MA confidence band). The MA confidence band is a very small improvement of the MAR confidence band obtained. However it is more difficult to calculate. This can be seen in Figures 2.4, 2.5 and 2.6, where one cannot distinguish between these two confidence bands. So in this work we just use the Frey minimum area confidence region (MA confidence region) and use Equation (2.7) to obtain the MAR confidence band. In Figures 2.4, 2.5 and 2.6, the Frey confidence bands have smaller area.

2.2.5 Distribution-free confidence bands

For distribution-free confidence bands, the most known method is the Kolmogorov [Birnbaum 52] statistic for small sample sizes and the Kolmogorov-Smirnov test for large sample sizes. Some other methods have also been suggested based on the weighted version of the Kolmogorov-Smirnov test [Anderson 52]. Owen [Owen 95] inverted the nonparametric likelihood test of uniformity introduced by Berk and Jones [Berk 78] to construct nonparametric likelihood

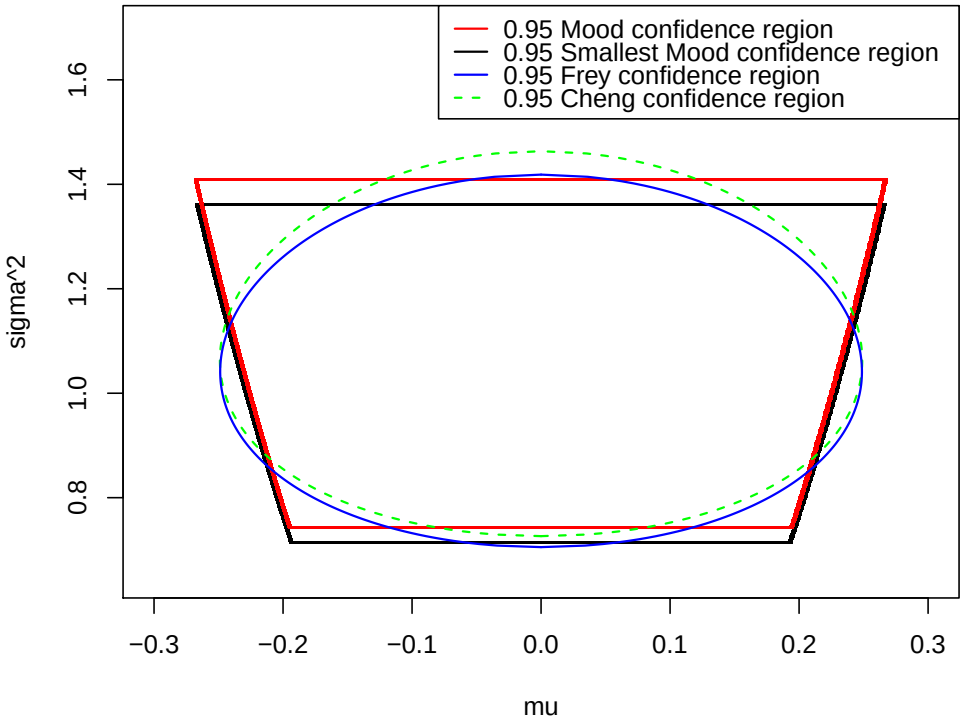

Figure 2.3: 0.95 confidence region for parameters of a normal distribution based on a sample set with size $n = 100$ and with $(\bar{X}, S) = (0, 1)$.

confidence bands for a distribution function. Nonparametric likelihood bands are narrower in the tails and wider in the center than Kolmogorov-Smirnov bands. They are asymmetric about the empirical cumulative distribution function. Frey [Frey 08] suggested another approach in which the upper and lower bounds of the confidence band are chosen to minimize a narrowness criterion. The optimal bands have a nice property: by choosing appropriate weights, one may obtain bands that are narrow in whatever region of the distribution is of interest. Other methods for construction of continuous confidence bands for parametric function have been proposed by Kanofsky and Srinivasan [Kanofsky 72] and Cheng and Iles [Cheng 83]. In Figures 2.7, 2.8 and 2.9, we have illustrated the Kolmogorov-Smirnov 0.95-confidence band for a sample set drawn from normal distributions. Meanwhile, we illustrated that if we do not know the sample set distribution, distribution-free confidence bands are everywhere wider than normal confidence bands. It is important to notice that even for large samples, like $n = 100$ in figure 2.9, the Kolmogorov-Smirnov band remains significantly wider than the parametric confidence bands.

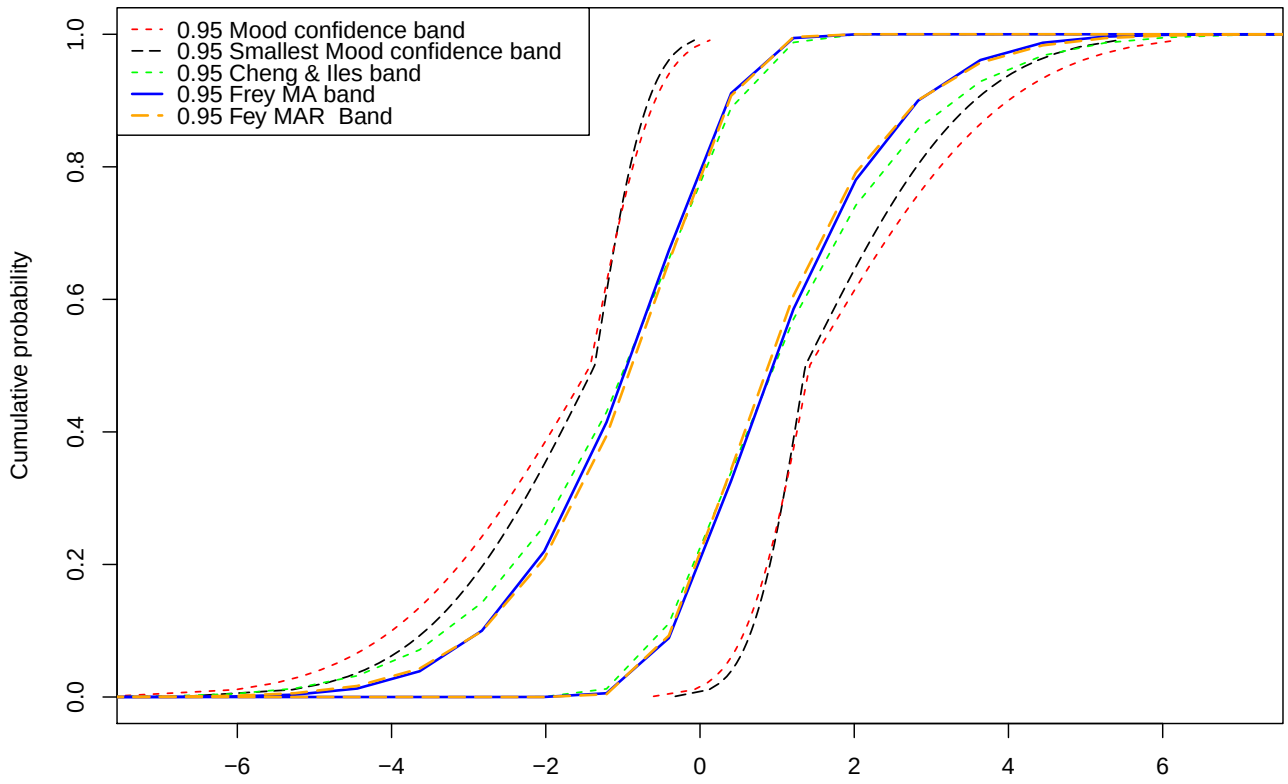

Figure 2.4: 0.95 confidence band for a normal distribution based on a sample set with size $n = 10$ and with $(\bar{X}, S) = (0, 1)$.

2.3 Tolerance interval

A tolerance interval is an interval that is guaranteed with a specified confidence level γ, to contain a specified proportion β of the population. Confidence bounds or limits are endpoints within which we expect to find a stated proportion of the population. As the sample set grows, a parameter's confidence interval decreases toward zero. In the same way, increasing the sample size leads the tolerance interval bounds to converge toward a fixed value. We name a $100\beta\%$ tolerance interval(region) with confidence level $100\gamma\%$, a β-content γ-coverage tolerance interval (region) and we denote it by $I^T_{\gamma,\beta}$.

Definition 13 *Let $X_1, \cdots, X_n$ denote a random sample from a continuous probability distribution and let $\mathbb{X} = (X_1, \cdots, X_n)$. A tolerance interval is an interval that is guaranteed, with a specified confidence level γ, to contain a specified proportion β of the population. The*

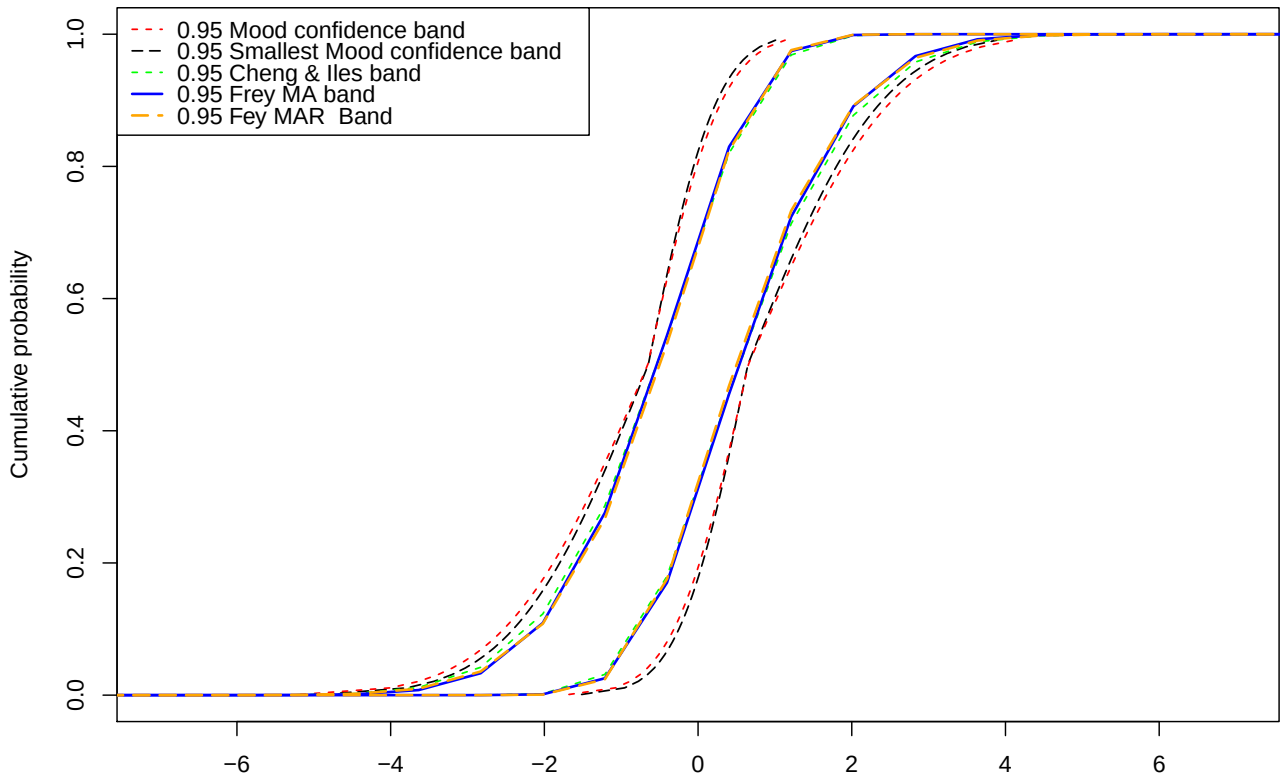

Figure 2.5: 0.95 confidence band for a normal distribution based on a sample set with size $n = 25$ and with $(\bar{X}, S) = (0, 1)$.

$I^T_{\gamma,\beta}$ sign is used to refer to a β-content γ-coverage tolerance interval [Krishnamoorthy 09]. Then, we have:

$$P_{\mathbb{X}}\Big(P(X \in I^T_{\gamma,\beta}|\mathbb{X}) \geq \beta\Big) = \gamma. \tag{2.9}$$

Suppose that we draw many independent groups of random samples from the distribution F. If one calculates the β-content γ-coverage tolerance interval from many of these groups of random samples, a γ fraction of tolerance intervals would, in the long run, contain at least a β proportion of F [Hahn 91].
Therefore one-sided tolerance intervals can be used as γ-level confidence intervals for an unknown β-quantile. In the same manner, two-sided tolerance intervals could serve as γ-level confidence intervals for an unknown β-inter-quantile. However we do not have the

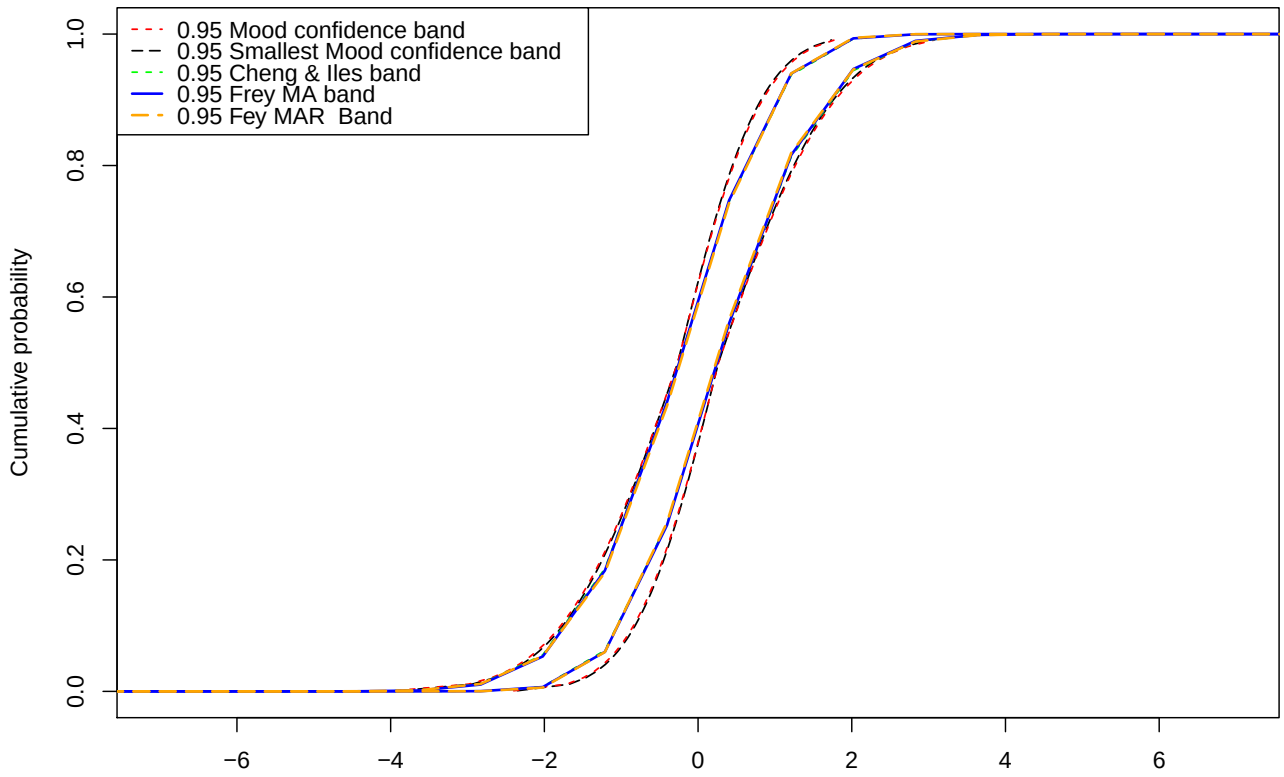

Figure 2.6: 0.95 confidence band for a normal distribution based on a sample set with size $n = 100$ and with $(\bar{X}, S) = (0, 1)$.

simultaneous condition given by (2.4).

$$P\Big(P(X \in I^T_{\gamma,\beta_i}) \geq \beta_i\Big) = \gamma, \forall \beta_i \in (0,1), i \in \{1, 2, \cdots, n\}, \tag{2.10}$$

but

$$P\Big(P(X \in I^T_{\gamma,\beta_1}) = \beta_1) \cap (P(X \in I^T_{\gamma,\beta_2}) = \beta_2)) \cap \cdots \cap (P(X \in I^T_{\gamma,\beta_n}) = \beta_n)\Big) \neq \gamma.$$

2.3.1 Tolerance interval for the Normal Distribution

When our sample set comes from a univariate normal distribution, the lower and upper tolerance bounds (x_l and x_u, respectively) are calculated by (2.11) and (2.12) where $\bar{X}$

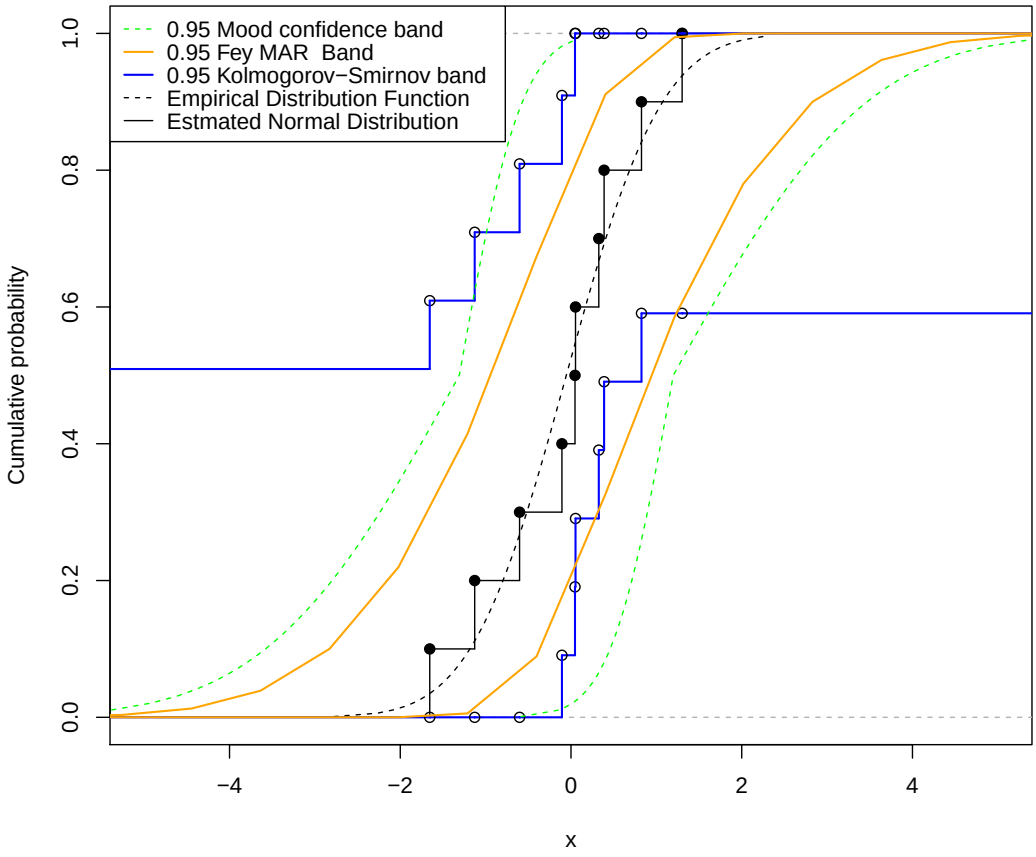

Figure 2.7: 0.95-Kolmogorov-Smirnov distribution free confidence band for a sample set with size $n = 10$ drawn from $\mathcal{N}(0, 1)$.

is the sample mean, S is the sample standard deviation, $\chi^2_{1-\gamma,n-1}$ represents the p-value of the chi-square distribution with $n-1$ degrees of freedom, and $Z^2_{1-\frac{1-\beta}{2}}$ is the square of the critical value of the standard normal distribution with probability $(1-\frac{1-\beta}{2})$ [Howe 69]. Hence, the boundaries of a β-content γ-coverage tolerance interval for a random sample of size n drawn from an unknown normal distribution are defined as follows:

$$x_l = \bar{X} - \mathbf{k}S, \;\; x_u = \bar{X} + \mathbf{k}S, \tag{2.11}$$

$$\mathbf{k} = \sqrt{\frac{(n-1)(1+\frac{1}{n})Z^2_{1-\frac{1-\beta}{2}}}{\chi^2_{1-\gamma,n-1}}}. \tag{2.12}$$

For more details on tolerance intervals see [Hahn 91, Krishnamoorthy 09].

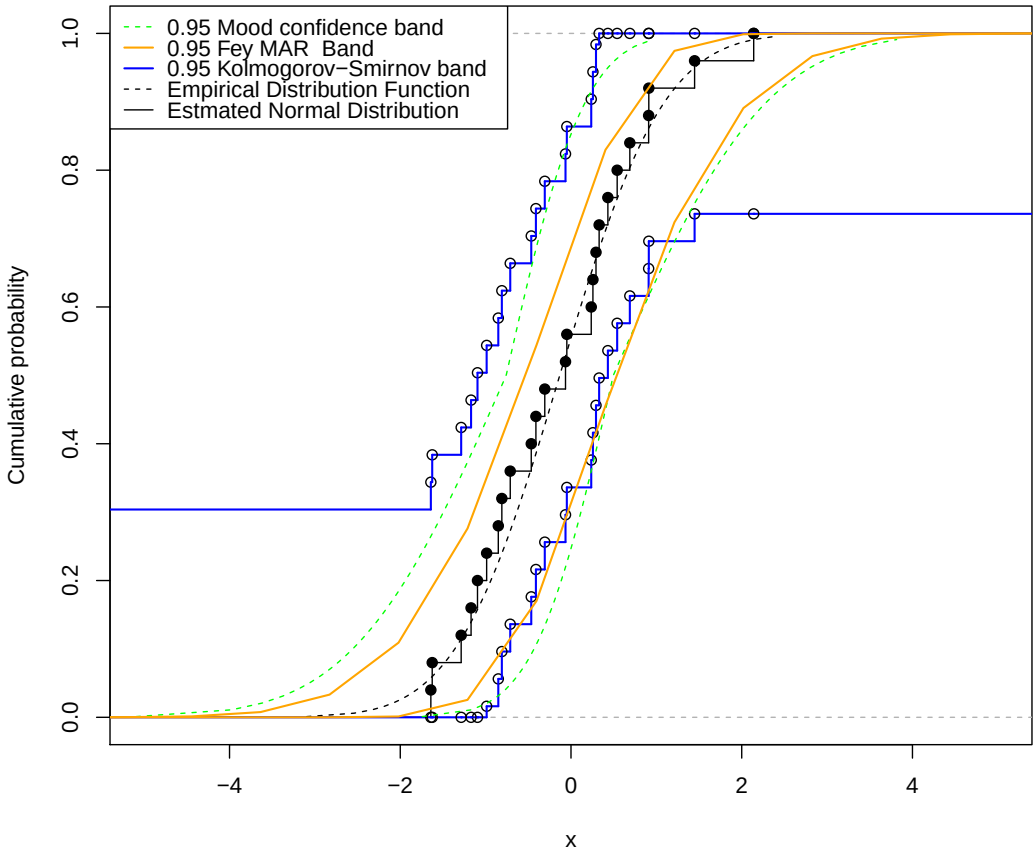

Figure 2.8: 0.95-Kolmogorov-Smirnov distribution free confidence band for a sample set with size $n = 25$ drawn from $\mathcal{N}(0, 1)$.

2.3.2 Distribution-free tolerance interval

Let $\{x_1, x_2, \cdots, x_n\}$ be n independent observations drawn from the continuous probability density function $f(x)$. A Distribution-free tolerance region is the region between two tolerance limits where the probability that this region contains a proportion β of the unknown probability distribution function is equal to γ. The mentioned tolerance limits are functions $L_1(x_1, x_2, \cdots, x_n) = x_r$, and $L_2(x_1, x_2, \cdots, x_n) = x_s$ constructed based on the order statistics of the observations:

$$\int_{x_s}^{x_r} f(x)dx \geq \beta, \tag{2.13}$$

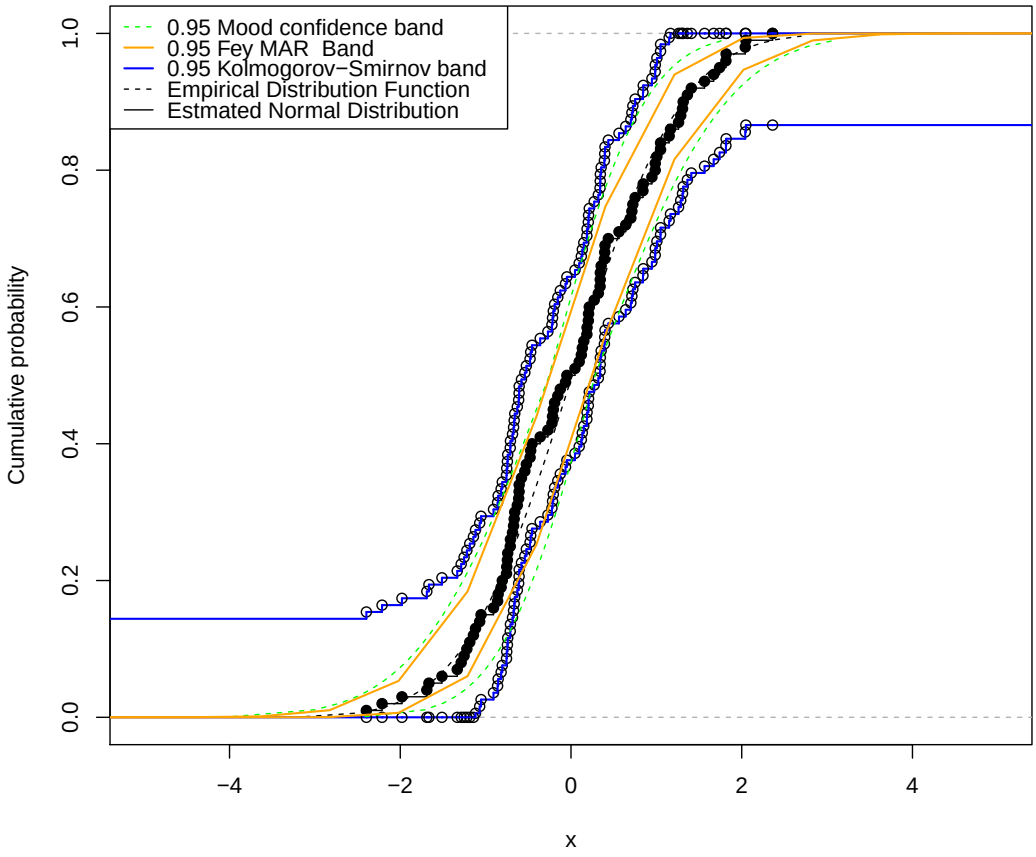

Figure 2.9: 0.95-Kolmogorov-Smirnov distribution free confidence band for a sample set with size $n = 100$ drawn from $\mathcal{N}(0, 1)$.

In order to find the Distribution-free β-content γ-coverage tolerance interval (region) of continuous random variable X, we have to find the smallest n and the order statistics x_r and x_s for which the probability that (2.13) holds is greater than or equal to γ. Equation (2.13) has a sampling distribution which was first defined by Wilks [Wilks 41] for a univariate random variable with symmetrical values of r and s. Wald [Wald 43] generalized the method to the multivariate case. The principle for finding a Distribution-free p-content $(1 - \alpha)$-coverage tolerance interval or region of continuous random variable X is based on order statistics. Having a univariate sample set, the distribution law, $f(p)$ of the p percent of the population of the universe included between the r and s order statistics is defined by

[Wald 43]:

$$f(p)dp = \frac{\Gamma(n+1)}{\Gamma(s-r)\Gamma(n-s+r-1)} p^{s-r-1}(1-p)^{n-s+r}dp. \tag{2.14}$$

In Wilks [Wilks 41] definition, $s = n + r - 1$. However in (2.14), as stated by [Wald 43], we can have any r and s such that $0 < r < s \leq n$. Note that because Distribution-free tolerance intervals are based on order statistics, the sample size required for a given Distribution-free tolerance interval may increase with the interval's confidence level (γ) or the interval's proportion β. For example, in order to have 95% 0.99-content tolerance interval between the first and last element of a sample set, using the formula in [Hanson 63], we need $n = 473$. For the calculation of the sample size requirement for tolerance intervals the reader can refer to [Hahn 91] and [Hanson 63].

2.3.3 Tolerance regions

The problem of normal tolerance intervals has been widely studied in the statistical literature. However the multivariate normal case has received less attention, especially when the number of variables is greater than two ($k > 2$). Wald [Wald 42] considered the tolerance region for large sample sizes. John [John 63] developed a theoretical framework and an approximation method to construct tolerance regions in a multivariate case and for finite sample size. Slotani [Slotani 64] also used approximations to build tolerance regions. However, the computations are considerable and he only gave a solution for the bivariate normal distribution. Chew [Chew 66] reviewed the result of John [John 63] and he also considered the other cases when the covariance matrix and/or the mean is known. Krishnamoorthy and Mathew [Krishnamoorthy 99] compared several tolerance region construction methods and use Monte Carlo simulation to show that all these approaches are inefficient with high dimension and high coverage probability ($p = 0.95$, $p = 0.99$). In the same paper two other approximation based approaches are proposed which give more satisfactory results. Krishnamoorthy and Mondala [Krishnamoorthya 06] suggested a new method which uses Monte Carlo methods combined with an approximation method to find the confidence factors. They also used Monte Carlo simulation to evaluate the accuracies of the tolerance and show that the new approach is very satisfactory, even for small samples. This approach is the most accurate solution to find the tolerance factors of the multivariate normal until now. More details can be found in [Krishnamoorthy 99] and [Krishnamoorthya 06].

Distribution-free tolerance regions for multivariate data were first addressed by Wald [Wald 43]. Then Tukey [Tukey 47],[Tukey 48] continued to study the construction of tolerance regions for the continuous and non-continuous random variables. He used Wald's principle to provide many new ways of using the samples of n to divide the range of the population into $n + 1$ blocks. For more detail, see [Tukey 47],[Tukey 48] and Murphy [Murphy 48].

2.4 Prediction interval

Let us now define a prediction interval and its associated possibility distribution. A prediction interval uses past observations to estimate an interval for future values. However other confidence intervals and credible intervals of parameters give an estimate for the unknown value of true population parameters.

Definition 14 *Let $X_1, X_2, \cdots, X_n$ be a random sample drawn from an arbitrary distribution, then interval $I_{1-\alpha}^{Prev} = [X_l, X_u]$ is a $100\%(1-\alpha)$ prediction interval such that:*

$$P\left(X_l \leq X_{n+1} \leq X_u\right) = 1 - \alpha.$$

$1-\alpha$ prediction intervals can be used as $(1-\alpha)$-level confidence intervals for the next observations and we denote them by $I_{1-\alpha}^{Prev}$. Equation (2.15) describes the predictive properties of prediction intervals.

$$P(X_{n+1} \in I_{1-\alpha}^{Prev}) = 1 - \alpha, \forall \alpha \in (0,1). \tag{2.15}$$

Suppose that we have many independent pairs of random samples. If in each pair we calculate the $(1-\alpha)$-prediction interval of the first sample and then see if the value(s) of the second sample are included in the computed $(1-\alpha)$-prediction interval of the first sample, $1-\alpha$ fraction of prediction intervals would, in the long run, contain the second sample value(s). Note that both the different pairs of samples and the observations within each sample must be independent [Hahn 91].

2.4.1 Prediction interval for the normal distribution

The prediction interval for the future observation from a normal distribution is given by [Hahn 69]:

$$\frac{x_{n+1} - \overline{X}_n}{S\sqrt{1+1/n}} \sim t_{n-1}, \tag{2.16}$$

$$I_{1-\alpha}^{Prev} = \left[\overline{X}_n - t_{(\frac{\alpha}{2},n-1)} S\sqrt{1+\frac{1}{n}}, \overline{X}_n + t_{(1-\frac{\alpha}{2},n-1)} S\sqrt{1+\frac{1}{n}}\right]. \tag{2.17}$$

Equation (2.17) gives a two-sided $1-\alpha$ prediction interval for the future observation x_{n+1} , where $\overline{X}_n$ represents the estimated mean from the n past observations, $t_{(1-\frac{\alpha}{2},n-1)}$ is the $100(\frac{1+p}{2})th$ quantile of Student's t-distribution with $n-1$ degrees of freedom.

For Distribution-free prediction intervals, the reader can find more information in [Hahn 91], [Konijn 87] and [Chakraborti 00].

2.4.2 Expectation Tolerance intervals

The tolerance intervals and regions mentioned before are β-content γ-confidence tolerance intervals. Another type of tolerance interval is the expectation tolerance interval.

Definition 15 *Let $X_1, \cdots, X_n$ denote a random sample from a continuous probability distribution and let $\mathbb{X} = (X_1, \cdots, X_n)$. A β-expectation tolerance interval is an interval which on average contains a specified proportion β of the population. The I_β^{EXT} notation, is used to refer to a β-expectation tolerance interval [Krishnamoorthy 09]. Then, we have:*

$$E_{\mathbb{X}}\Big(P(X \in I_\beta^{EXT}|\mathbb{X})\Big) = \beta. \tag{2.18}$$

An expectation tolerance interval or region is such that its average content is β. Paulson [Paulson 43] showed that the interval $[L(\mathbb{X}), U(\mathbb{X})]$ which based on a random sample $\mathbb{X}$, is a β-prediction interval for observing the next observation of the random variable X is also a β-expectation tolerance interval.

$$P_{\mathbb{X},X}(X \in I_\beta^{EXT}) = \beta. \tag{2.19}$$

2.5 Discussion

We have seen three types of intervals and their statistical interpretations. Each interval has its own application. Confidence bands are more suitable for making statements about the whole distribution. Tolerance intervals focus on the probability that one interval contains at least a desired proportion of the unknown population. Prediction intervals are suited for next future observations or mean coverage behaviors.

Example 1 ***Tolerance intervals :*** *suppose that we want to access the air lead level in a laboratory. We can see that the log transform of the sample data fits a normal distribution, so we will compute a $\gamma = 0.9$, $\beta = 0.95$ one-sided upper normal tolerance interval for the log transformed sample. Now if the obtained tolerance interval does not exceed the Occupation Exposure Limit (OEL), the laboratory is considered to be safe [Krishnamoorthy 09]. Tolerance intervals have industrial applications like quality control, environmental monitoring, industrial hygiene, exposure data analysis, etc. For more examples, see [Gibbons 01] and [Gibbons 94].*

Example 2 ***Prediction intervals :*** *an automobile client may wish to know, based on a sample set of five similar cars, an interval that with a high degree of confidence will cover the gasoline millage that the new automobile will obtain under specified driving conditions [Hahn 91]. In this case we assume that the five automobiles and the new one are random observations from the same population.*

Example 3 ***Confidence band :*** *suppose that we are a transport company and we transport pieces with different weights. We have a sample of the 50 past transports in the current year and we wish to build the next year's pricing strategy. For this purpose, we need to know the maximum and minimum weights of* $0.5, 0.75, 0.95$ *and* 0.99 *fraction of our future commands. We will use these intervals to simultaneously build the future sales strategy. In this case we will find a high* $\gamma = 0.99$ *confidence band for the unknown distribution. Then we use this band to find simultaneous confidence intervals on population quantiles. Then we have a probability of* $1-\gamma = 0.01$ *that all these intervals simultaneously do not cover the real fraction of weights. So we have a probability of* 0.99 *to have a correct sale strategy. However if we used two-sided* 0.95 *coverage* $0.5, 0.75, 0.95$ *and* 0.99 *-content tolerance intervals, we would have a chance of* 0.01 *of making a mistake for each interval and that at least one not covering interval is stronger than* 0.01*.*

We can use (2.4), (2.10) and (2.15) to deduce the following properties:

Proposition 3

$$\forall\gamma, \forall\beta,\ I^c_\beta \geq I^T_{\gamma,\beta}. \tag{2.20}$$

All the simultaneous γ*-confidence intervals of* β*-interquartile for an unknown distribution are wider than their corresponding* β*-content* γ*-coverage tolerance intervals.*

Proposition 4

$$\forall n \geq 5, \gamma \geq 0.75, \forall\beta,\ I^c_\beta \geq I^T_{\gamma,\beta} \geq I^{Prev}_\beta. \tag{2.21}$$

For any random sample larger than 5*, if we fix* γ *and* β*, then the* β*-inter-quantile of confidence band is equal to or larger than the corresponding* β*-content tolerance intervals and* β*-prediction intervals the smallest ones.*

These two properties can be easily verified by comparing numerical values of the aforementioned confidence bands, tolerance and prediction formula. The reader can also refer to tables and equations listed in [Hahn 91] and [Krishnamoorthy 09].

2.6 Conclusion

This chapter presents the definition and the interpretation of confidence bands, tolerance intervals and prediction intervals. We have seen that if we want a γ-confidence interval on a β-inter-quantile, we can infer them from tolerance intervals or confidence bands but those provided by confidence bands are always wider than their corresponding tolerance intervals. We have also noticed that prediction intervals are expectation tolerance intervals. So, as opposed to γ-coverage β-content tolerance intervals that with confidence level γ, contain β proportion of the underlying distribution, β-prediction intervals (or β-expectation tolerance intervals) are intervals that, on average, contain a proportion β of the underlying population. The next chapter shows how these intervals can be encoded by possibility distributions.

Chapter 3

Encoding a family of probability distribution

Contents

For a given sample set, there are already different possibility distributions that encode a family of probability distributions that may have generated our sample set. Apart from our recent study [Ghasemi Hamed 12b], almost all the existing methods are based on parametric and distribution-free confidence bands. In this work, we look at these new possibility distributions. These distributions encode different kinds of uncertainties that have not been treated before. They encode statistical tolerance and prediction intervals (regions). We also proposed a possibility distribution encoding the confidence band of the normal distribution which improves on the existing one for all sample sizes. In this work we keep the idea of building possibility distributions based on intervals which are among the smallest intervals for small sample sizes. We also discuss the properties of the mentioned possibility distributions. This chapter is a detailed version of our work in [Ghasemi Hamed 12b] and [Bounhas 13]. Our contributions are some comparative figures or concluding propositions not stated in the original paper.

In the previous chapter, we reviewed some methods for constructing confidence bands for the normal distribution and for constructing distribution-free confidence bands (γ-C distribution). Here we propose a possibility distribution for a sample set drawn from an

unknown normal distribution based on Frey's [Frey 09] confidence band which improves the existing possibility distribution proposed by Aregui et al. [Aregui 07a] for all sample sizes. We also introduced a possibility distribution which encodes tolerance intervals, called the γ-CTP distribution [Ghasemi Hamed 12b]. The proposed possibility distribution uses tolerance intervals to build the maximal specific possibility distribution that bounds each population quantile of the true distribution (with a fixed confidence level) that might have generated our sample set. The distribution obtained will bound each confidence interval of inter-quantiles independently. This latter is different from a possibility distribution encoding a confidence band, because a possibility distribution encoding a confidence band will simultaneously bound all population quantiles of the true distribution (with a fixed confidence level) that might have generated our sample set. Finally, we consider possibility distributions encoding prediction intervals (prediction possibility distribution). In this case, each α-cut will contain the next observation with a confidence level equal to $1-\alpha$. Each of the proposed possibility distributions encodes a different kind of uncertainty that is not expressed by the other ones. We show that a γ-confidence distribution is always less specific than a γ-CTP distribution which is itself less specific than the prediction possibility distribution. This is due to the fact that the distribution's properties are less and less strong. Note that the confidence level is usually chosen by the domain expert. This section is structured as follows: we begin with a review of possibility distribution encoding confidence bands and their relationship with confidence regions. In this section we introduce a method which improves existing possibility distributions. Next we see how to encode tolerance intervals and prediction intervals by possibility distributions. Finally, we end with a discussion of the mentioned possibility distributions and some illustrations.

3.1 Possibility distribution encoding confidence bands

We saw in 2.2.1 that a confidence band is an interval defined for each value x of the random variable X such that for a repeated sampling, the frequency of $F(x)$ located inside the interval $[L(x), U(x)]$ for all the values of X tends to the confidence coefficient γ. It can also be used to infer confidence intervals of the quantile function (see 2.2.1). In other words, the confidence band simultaneously gives confidence intervals for all quantiles. Therefore such confidence intervals derived from confidence bands are *Simultaneous confidence Intervals (SCI) for all population quantiles*. We tooked advantage of this property to derive simultaneous γ-confidence intervals for β-content inter-quantiles of the unknown cdf $F(\cdot)$ and denoted them by I^C_β.

By using proposition (1) and tables of confidence band stated in the statistic literature [Anderson 52, Kanofsky 72, Birnbaum 52, Cheng 83, Frey 09, Frey 08], we can encode simultaneous γ-confidence intervals for β-content inter-quantiles I^C_β, of an unknown cdf $F(\cdot)$ by a possibility distribution represented by π^C_γ:

$$\pi^C_\gamma(x) = 1 - \underset{x \in I^C_{1-\alpha}}{max}(\alpha) \text{ where } A_\alpha = I^C_\beta, \beta = 1-\alpha. \tag{3.1}$$

By construction, the obtained distribution has the following property:

Proposition 5 *Let π_γ^C be a possibility distribution obtained by Equation (3.1). We have:*

$$P(\forall \alpha \in (0,1),\ P(X \in A_\alpha) \geq 1-\alpha) \geq \gamma.$$

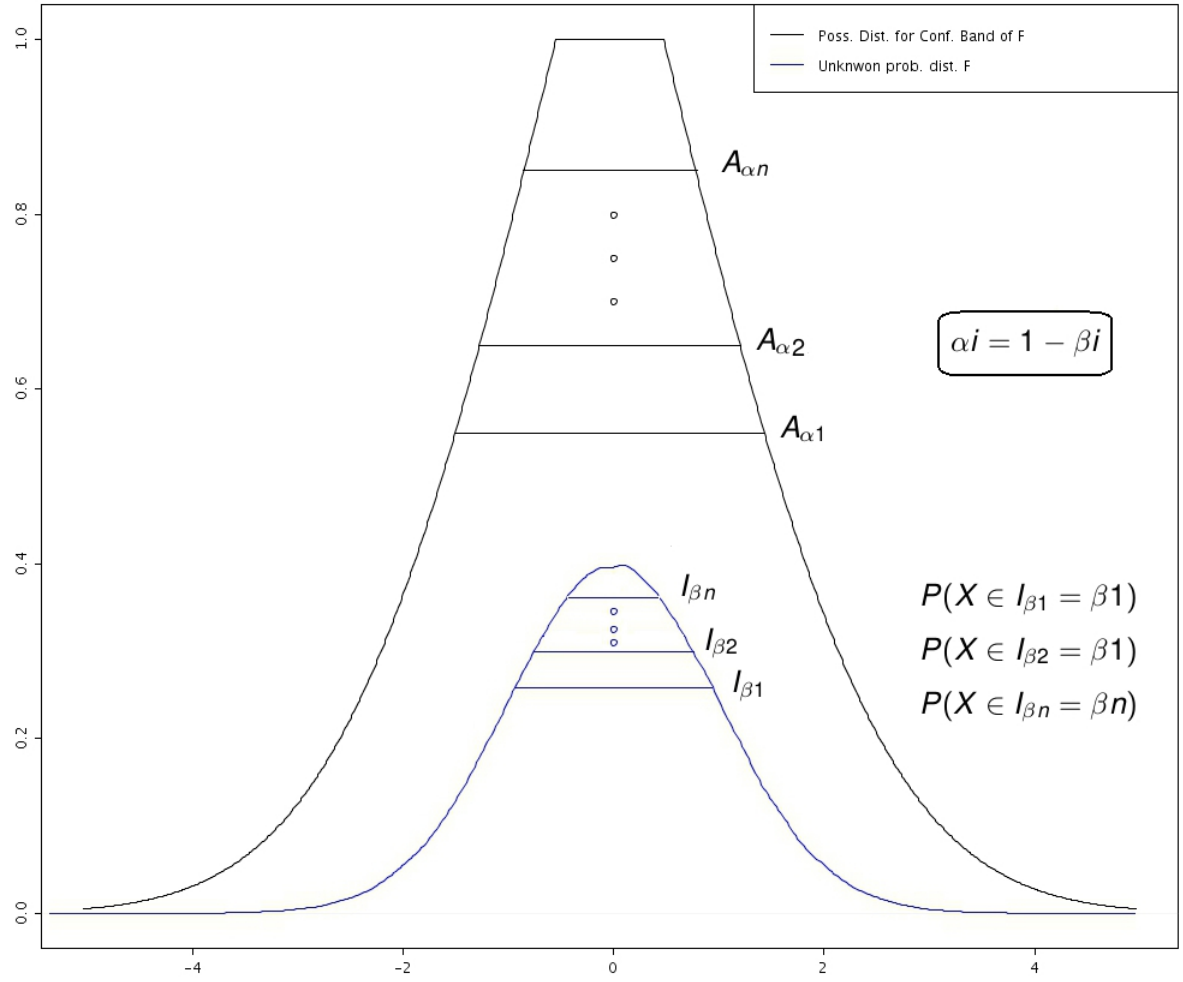

Figure 3.1: Comparing inter-quantiles of $\mathcal{N}(0,1)$ with its 0.95-Confidence distribution based on a sample set with $(\mu, \sigma^2) = (\bar{X}, S^2) = (0,1)$.

Here, we restate an equation similar to Equation (2.4), but expressed for confidence-possibility distributions.

$$P((P(X \in A_{\alpha_1}) = \alpha_1) \cap (P(X \in A_{\alpha_2}) = \alpha_2)) \cap \cdots \cap (P(X \in A_{\alpha_n}) = \alpha_n))) = \gamma, \quad (3.2)$$

where $\alpha_i = 1 - \beta_i$. Figure 3.1 shows this concept graphically. The blue probability distribution denotes the unknown probability distribution which has generated the sample set

and the black possibility distribution is the corresponding 0.95-Frey Confidence distribution. In order to ease the comparison of inter-quantiles and their corresponding α-cut we suppose that the sample set which has $n = 10$ comes from $\mathcal{N}(\mu, \sigma^2)$ and $(\mu, \sigma^2) = (\bar{X}, S^2) = (0, 1)$.

3.1.1 Possibility distribution encoding normal confidence bands

As stated in section 2.2.1, we can use the confidence region for the parameter of a probability distribution to infer its confidence band. Thus a possibility distribution that encodes a $1-\alpha$ confidence region for the parameters of a normal distribution has the same properties as a possibility distribution encoding a $1-\alpha$ normal confidence band. This is stated by the following proposition:

Proposition 6 *Suppose that we have a sample set of n observations from a normal distribution and let $\overline{X}, S^2$ and $\mathcal{R}^{1-\alpha}$ respectively denote the sample mean, the sample variance and the $(1-\alpha)$-confidence region of the parameters of the normal distribution that may have generated our random sample. Let also, $\mathcal{F}$ represent the family of normal distribution that have their parameters inside region $\mathcal{R}^{1-\alpha}$.*

The most specific possibility distribution $\pi_{(n,\overline{X},S)}$ which encodes the family $\mathcal{F}$ has the same statistical properties as the most specific possibility distribution which encodes any $1-\alpha$-normal confidence for the mentioned sample.

Proof: We know that encoding the family $\mathcal{F}$ is similar to encoding the $1-\alpha$-confidence band resulted by $\mathcal{R}^{1-\alpha}$. So encoding the family $\mathcal{F}$ leads to encode a confidence band. It is obvious that two possibility distributions which encode two distinct normal confidence bands having the same confidence level $1-\alpha$ share the same statistical properties ■

This most specific possibility distribution encoding the family $\mathcal{F}$ is constructed by the formula below. Let $\Lambda = \{\pi | \pi = Tr(F), F \in \mathcal{F}\}$ be the set of possibility distributions obtained by applying the probability-possibility transformation $Tr(\cdot)$ (described by Equation (1.8)) to each probability distribution in $\mathcal{F}$. The possibility distribution defined by

$$\pi_{(n,\overline{X},S)}(x) = \sup\{\pi(x) | \pi \in \Lambda\}$$

encodes all the family $\mathcal{F}$ and has the following definition:

$$\pi_{(n,\overline{X},S)}(x) = \begin{cases} 1, & \text{if } x \in [\mu_{min}, \mu_{max}] \\ 2 * \mathcal{G}(x, \mu_{min}, \sigma^2_{max}), & \text{if } x < \mu_{min} \\ 2 * \mathcal{G}(2 * \mu_{max} - x, \mu_{max}, \sigma^2_{max}), & \text{if } x > \mu_{max} \end{cases} \tag{3.3}$$

where μ_{min}, μ_{max} and σ^2_{max} are respectively the lower and the upper bounds of the mean confidence interval, and the upper bound of the variance confidence interval associated to the confidence region found by (2.8). Moreover, $\mathcal{G}(x, \mu, \sigma^2)$ is the cumulative distribution function of $\mathcal{N}(\mu, \sigma^2)$.

Cheng and Iles [Cheng 83] and Kanofsky [Kanofsky 72] used this approach to infer the confidence band of the normal distribution. Aregui et al. [Aregui 07b], proposed to construct possibility distributions for a sample set drawn from a known parametric family with an unknown parameter vector. Their possibility distribution encoded the Cheng et al. [Cheng 83] confidence band. Aregui et al. [Aregui 07a] proposed a similar possibility distribution which encoded the "Smallest Mood exact" confidence region for parameters of the normal distribution. The "Smallest Mood exact" region contains exactly the desired confidence level and it was the the second smallest confidence region (after the "likelihood-ratio test") in [Arnold 98]. This region is easy to obtain and is particularly useful for small sample sizes.

Frey proposed the minimum-area confidence region and the minimum area based confidence band for the normal distribution. She showed that her minimum area confidence band improves on other bands for all sample sizes [Frey 09]. In the same way we propose a possibility distribution which encodes the Frey confidence band. In Figures (3.3,3.4) we compare our possibility distribution named "0.95 Frey C.P.D." (0.95 Frey Confidence Possibility Distribution) which is displayed in blue, with the Mood based and Smallest Mood based confidence possibility distribution. The Mood based and the Smallest Mood based confidence bands are obtained by Equation (2.7). We have seen that Frey's normal confidence band improves on the confidence band resulted by the "Smallest Mood exact" region and the situation is the same for the encoding possibility distributions.

3.1.2 Possibility distribution encoding distribution-free confidence bands

Masson et al. [Masson 06], suggested simultaneous confidence intervals of the the multinomial distribution to build possibility distributions. In another paper, Aregui et al. [Aregui 07b] proposed the Kolmogorov confidence band [Birnbaum 52] to construct predictive belief functions [Smets 90b] for sample set drawn from an unknown distribution. Thus, we propose use of Frey's band to construct the possibility distribution, since it allows us to have narrower α-cuts for the α's of interest.

3.2 Possibility distribution encoding tolerance interval

Tolerance intervals were defined in section 2.3. A β-content γ-coverage tolerance interval is denoted here by $I^T_{\gamma,\beta}$. Having a sample set which comes from a cdf $F(\cdot)$ with unknown parameters and for a given confidence level γ, we encode all the β-content γ-coverage tolerance intervals of $F(\cdot)$, $\forall\beta \in (0,1)$, by a possibility distribution and we name it "γ-confidence tolerance possibility distribution" (γ-CTP distribution represented by π^{CTP}_{γ}). When we do not know the distribution of the sample set, we can use β-content γ-coverage distribution-free tolerance intervals ,$\forall\beta \in (0,1)$, of the unknown probability distribution in order to build distribution-free γ-Confidence Tolerance Possibility (γ-DFCTP distribution

represented by π_γ^{DFCTP}) distribution. The possibility distributions π_γ^{CTP} and π_γ^{DFCTP} will have, by construction the following property:

Proposition 7 *Let π_γ^{CTP} (or π_γ^{DFCTP}) be a possibility distribution that encodes tolerance intervals. We have:*

$$\forall \alpha \in (0,1),\ P(P(X \in A_\alpha) \geq 1-\alpha) \geq \gamma, \text{where } A_\alpha = I^T_{\gamma,\beta}, \beta = 1-\alpha. \tag{3.4}$$

Note that it may also be interesting to fix the proportion β and make the confidence coefficient vary, $\gamma \in (0,1)$, to have a β-content tolerance possibility distribution.

Equation (3.4) is the same as Equation (2.10) but is now stated for CTP distributions or Distribution Free Confidence Tolerance Possibility Distribution (DFCTP distribution)s and Figure 3.2 shows this concept graphically. In Figure 3.2, the blue probability distribution denotes the unknown probability distribution which has generated the sample set and the black possibility distribution is the corresponding 0.95-CTP distribution. In order to ease the comparison of inter-quantiles and their corresponding α-cut we supposed that the sample set which has $n = 10$ comes from $\mathcal{N}(\mu, \sigma^2)$ and $(\mu, \sigma^2) = (\bar{X}, S^2) = (0, 1)$.

Possibility distribution encoding tolerance interval for the normal distribution

When our sample set comes from a univariate normal distribution, the lower and upper tolerance bounds (x_l and x_u respectively) are calculated by formulas (2.11) and (2.12). By using proposition (1), we can find the boundaries of the $(1-\alpha)$-cut $A_{1-\alpha} = [x_l, x_u]$ of the possibility distribution which are calculated by (2.11), then we obtain the possibility distribution π_γ^{CTP} as computed below, where $\Phi(\cdot)$ is the cdf of the standard normal distribution.

$$\pi_\gamma^{CTP}(x) = 2\left(1 - \Phi\left(\sqrt{\frac{\chi^2_{(1-\gamma,n-1)}(\frac{x-\bar{X}}{S})^2}{(n-1)(1+\frac{1}{n})}}\right)\right). \tag{3.5}$$

Possibility distribution encoding distribution-free tolerance interval

The construction of possibility distribution based on distribution-free tolerance intervals (region) raises some problems, because for a given sample set there are many ways to choose the r and s order statistics. If we choose them symmetrically like in the Wilks method [Wilks 41] ($r = n - s + 1$), then the possibility distribution which encodes these intervals does not guarantee that its α-cuts include the mode and the α-cuts are neither the smallest ones. In fact, for any symmetric unimodal distribution, if we choose r and s order statistics in a symmetrical way, we will have tolerance intervals which are also the smallest possible ones and also include the mode of the distribution (see proposition (1)). Thus the Distribution-Free γ-Confidence Tolerance Possibility (π_γ^{DFCTP}) distribution is constructed by the following equation where x_r and x_s are the limits for the distribution-free $I^T_{\gamma,\beta}$ of our sample set.

$$\pi_\gamma^{DFCTP}(x) = 1 - \max_{x \in I^T_{\gamma,1-\alpha}} (\alpha), \text{where } A_\alpha = I^T_{\gamma,\beta} = [x_r, x_s], \beta = 1-\alpha$$

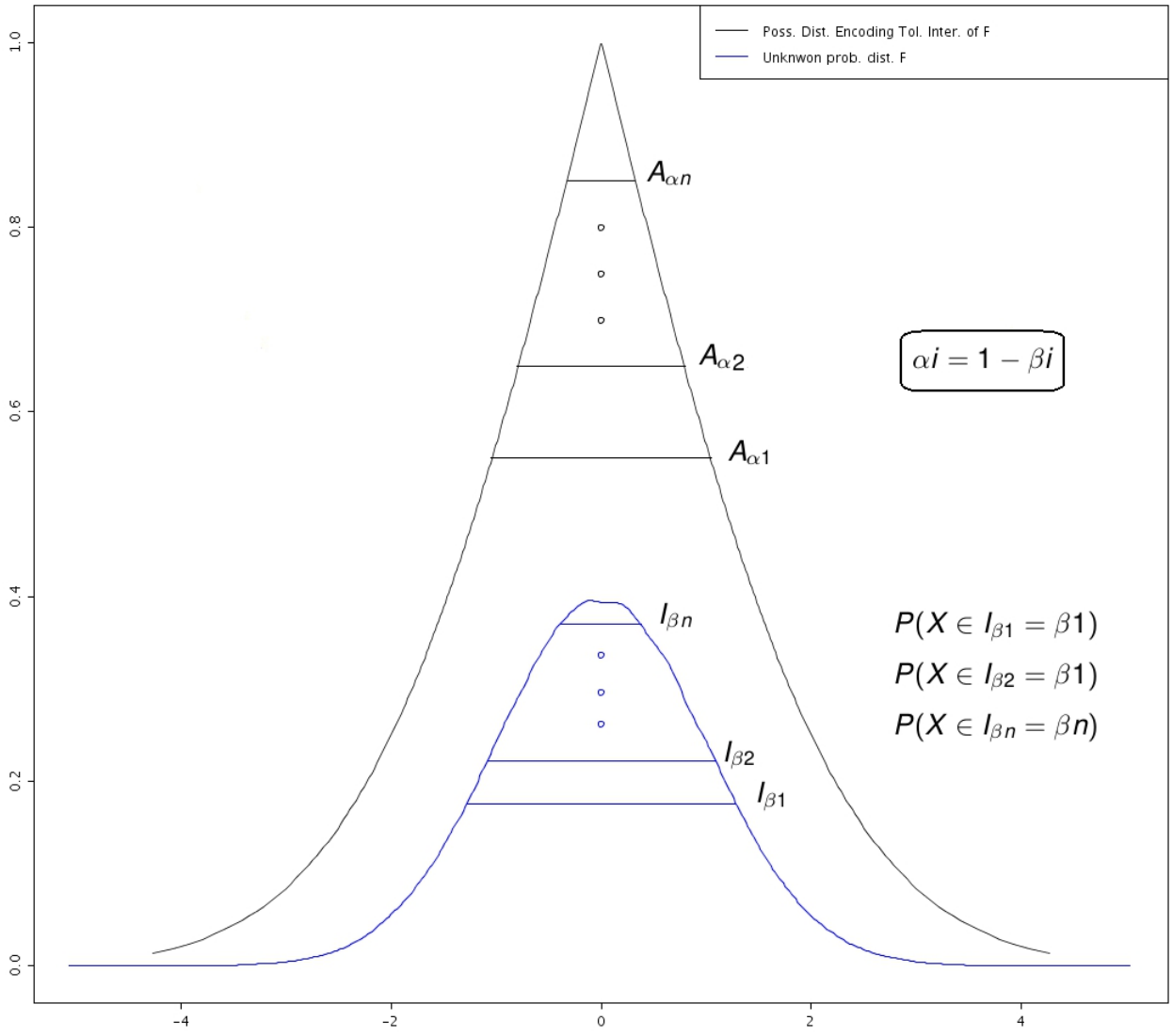

Figure 3.2: Comparing inter-quantiles of $\mathcal{N}(0,1)$ with its 0.95-CTP distribution based on a sample set with $(\mu, \sigma^2) = (\bar{X}, S^2) = (0,1)$.

3.3 Possibility distribution encoding prediction intervals

A prediction interval, defined in section 2.4, uses past observations to estimate an interval for what the future values will be and we denote a $1-\alpha$ prediction interval by I_β^{Prev} where $\beta = 1-\alpha$. By using Proposition (1) and Equation (2.17), we can infer a prediction possibility (π^{Prev}) distribution for a sample set which comes from a normal distribution with an unknown mean and variance. π^{Prev} is computed as below, where $T_{n-1}(\cdot)$ is the cdf of the Student t distribution with $n-1$ degree of freedom. Equation (3.6) describes how to

compute a prediction possibility distribution and Proposition 8 shows its α-cut properties.

$$\pi^{Prev}(x) = 2\Big(1 - T_{n-1}\Big(\Big|\frac{X_{n+1} - \overline{X}_n}{S\sqrt{1+1/n}}\Big|\Big)\Big). \tag{3.6}$$

By construction, the obtained distribution has the following property:

Proposition 8 *Let π^{prev} be a possibility distribution that encodes prediction intervals using equation (3.6) built from a random sample set $\mathbf{X} = \{X_1, \ldots, X_n\}$ we have:*

$$\forall \alpha \in (0,1),\ P(X_{n+1} \in A_\alpha) \geq 1 - \alpha, \text{where } A_\alpha = I_\beta^{Prev}, \beta = 1 - \alpha.$$

3.4 Discussion and Illustrations

We have seen three different types of intervals and their encoding possibility distributions. The most common approach is to choose the possibility distribution which is encoded by confidence bands. However, depending on the application, we might be interested to infer other possibility distributions than the one that encodes conventional Simultaneous Confidence Intervals (SCI)s. Section 2.5 discusses different applications of the different intervals encoded by the mentioned possibility distributions. Figure (3.3) shows the $\pi^C_{0.95}$ for a sample set of size 10 with sample mean and sample variance respectively equal to 0 and 1. Figure (3.3) represents the same concept for $n = 25$. This figure illustrates the final remark in Section 3.1.1. Indeed, we can see that our possibility distribution is more informative than the Aregui et al. possibility distribution.

In Figure (3.6) the blue color is used to represent π^{Prev} for different sample sets drawn from the normal distribution, all having the same sample parameters, $(\overline{X}, S) = (0, 1)$ but different sample sizes. The green distribution represents the probability-possibility transformation of $\mathcal{N}(0, 1)$.

In Figure (3.5) we used the previous settings for the $\pi^{CTP}_{0.95}$. Note that, for $n \geq 100$, the tolerance interval is approximately the same as the maximum likelihood estimated distribution. In Figure 3.7, the blue curves represents the $\pi^{DFCTP}_{0.95}$ for a sample set of size 450, drawn from $\mathcal{N}(0, 1)$ and the green distribution represents the probability-possibility transformation for $\mathcal{N}(0, 1)$. In Figure (3.8), we used two different sample sets with $n = 194$ to build two different $\pi^{DFCTP}_{0.9}$. In this example, in order to reduce the required sample size, we restricted the largest β to 0.98.

Figures (3.4, 3.4 and 3.4), represent and compare the three possibility distributions. It is easy to note that the 0.95-Frey confidence possibility distribution, shown by the black color mentioned above, always provide wider intervals. Then comes the 0.95-CTP distribution and the smallest one is the 0.95 prediction possibility distribution. The blue color is used to represent π^{Prev} for different sample sets drawn from the normal distribution, all having the same sample parameters, $(\overline{X}, S) = (0, 1)$ but different sample sizes. The green distribution represents the probability-possibility transformation of $\mathcal{N}(0, 1)$. Finally, we can deduce from Proposition 4 that:

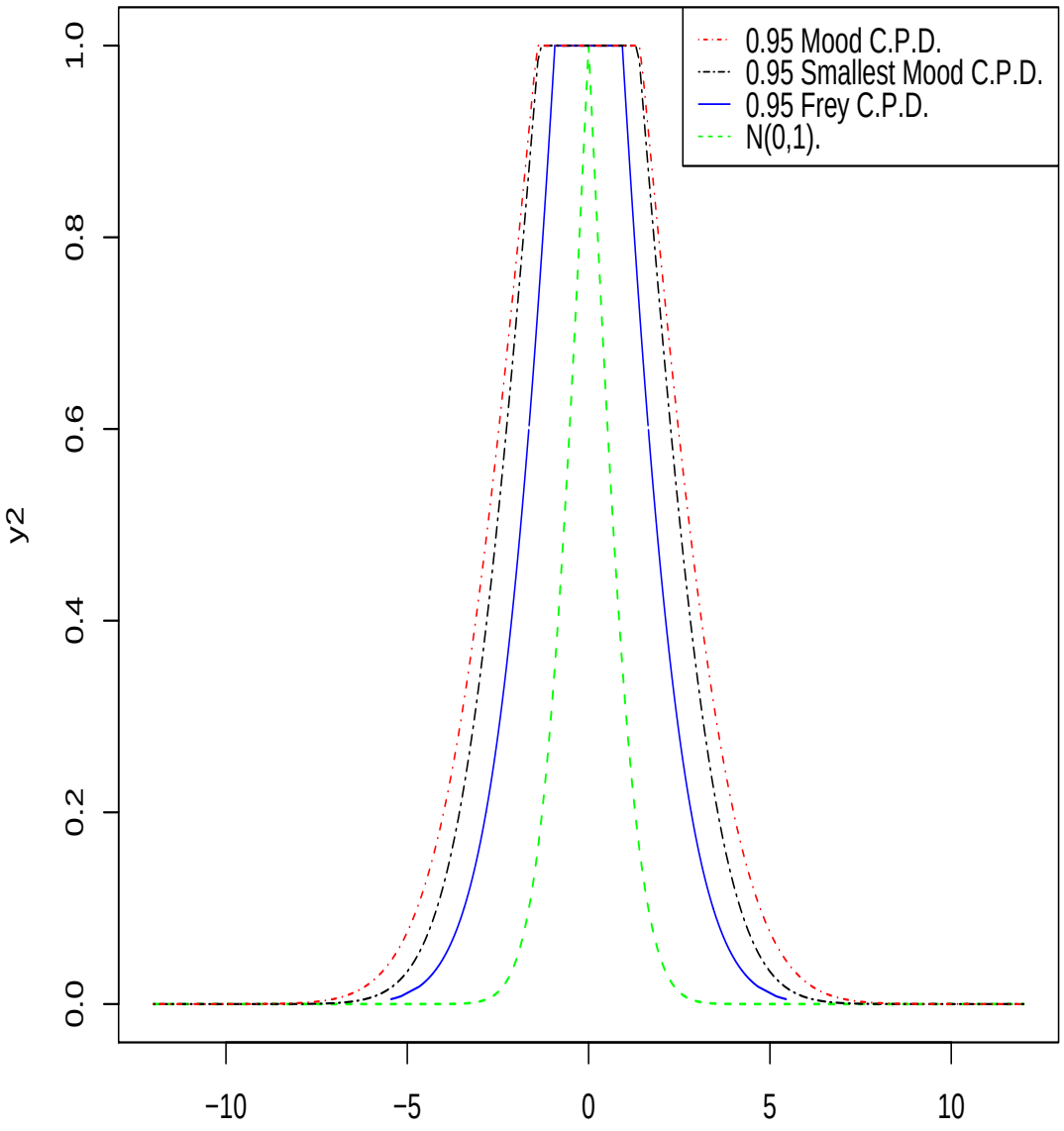

Figure 3.3: Possibility distribution encoding normal confidence band for a sample set of size 10 having $(\bar{X}, S) = (0, 1)$.

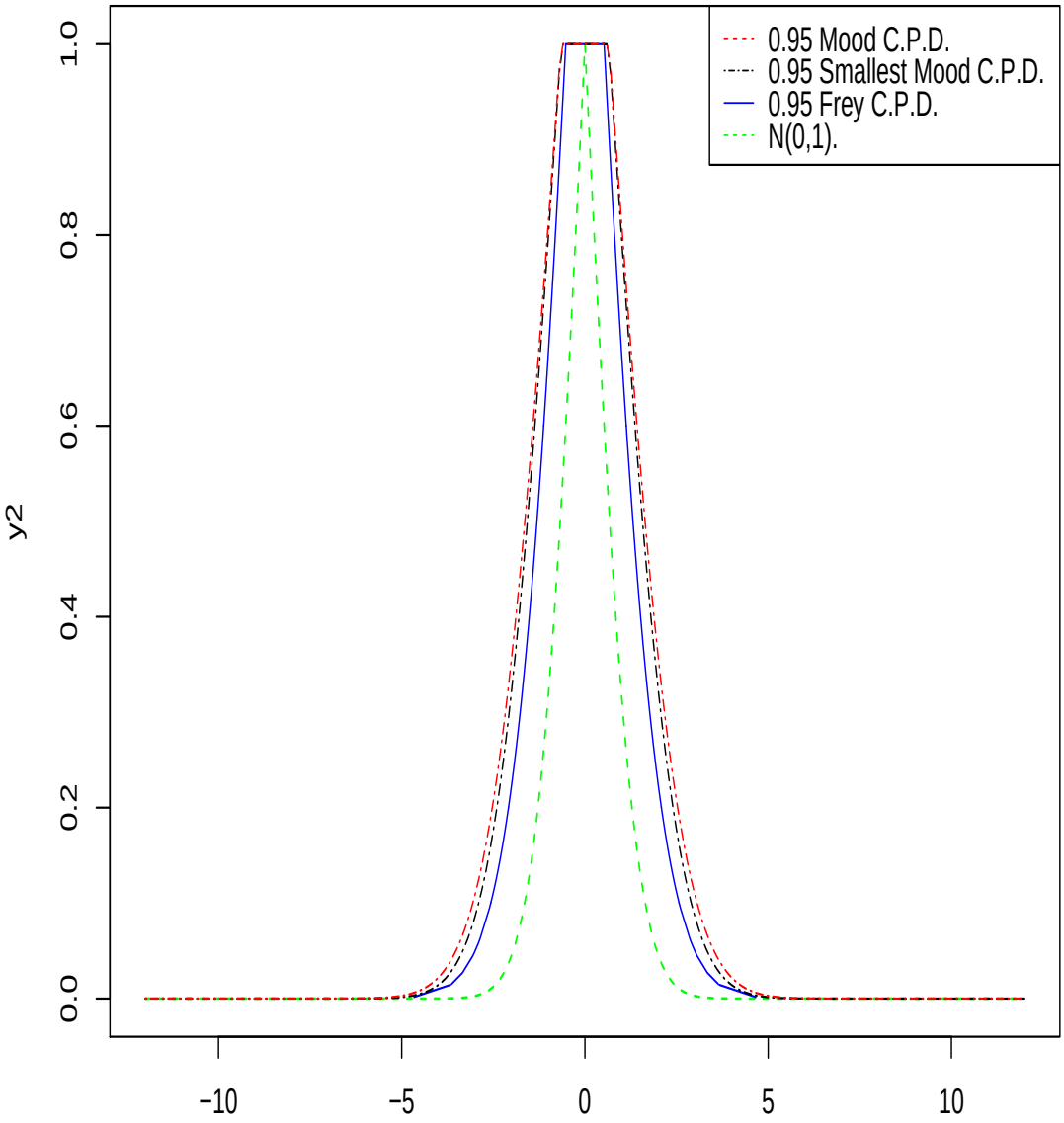

Figure 3.4: Possibility distribution encoding normal confidence band for a sample set of size 25 having $(\bar{X}, S) = (0, 1)$.

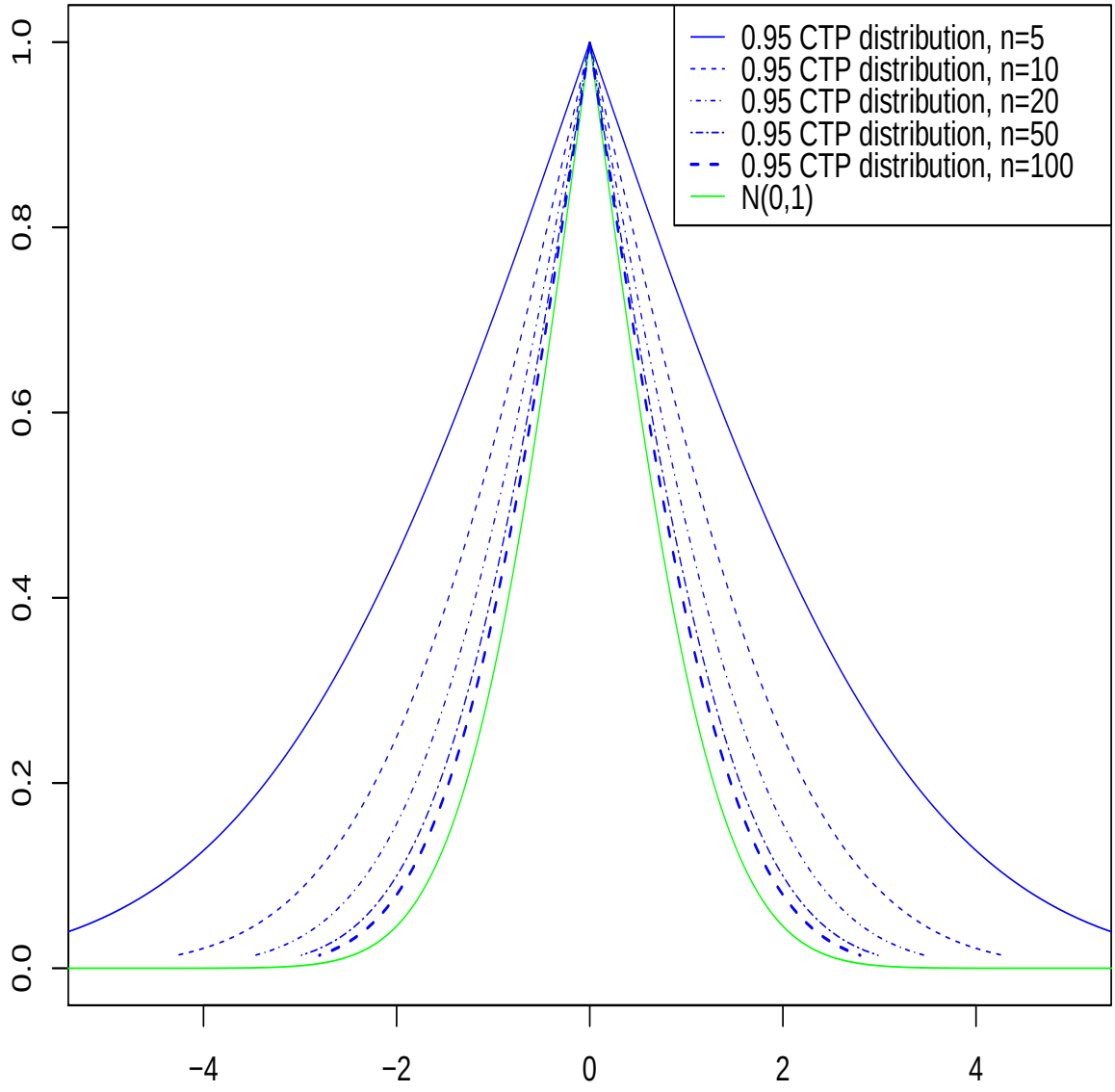

Figure 3.5: 0.95-confidence tolerance possibility distribution for different sample sizes having $(\overline{X}, S) = (0, 1)$.

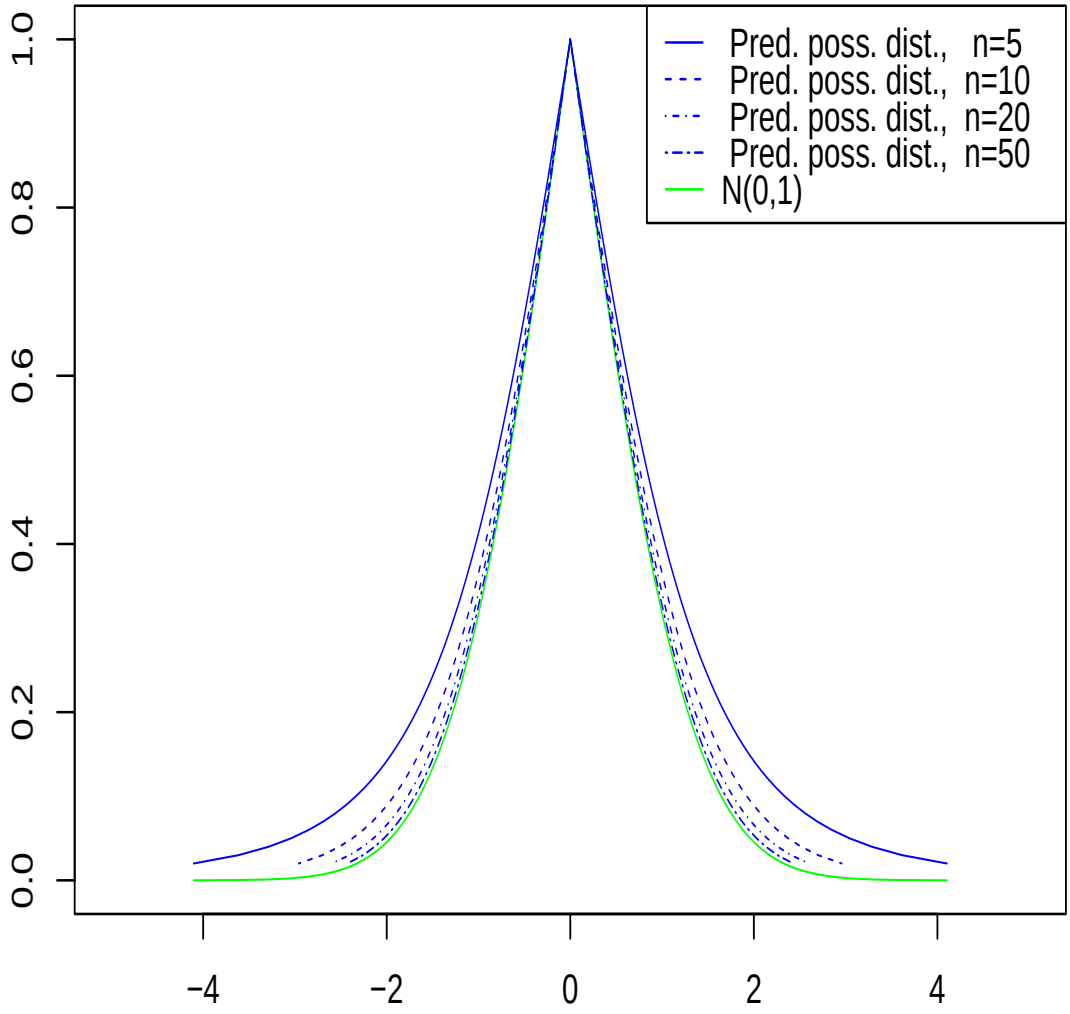

Figure 3.6: 0.95-confidence prevision possibility distribution for different sample sizes having $(\overline{X}, S) = (0, 1)$.

Proposition 9

$$\forall n \geq 5, \gamma \geq 0.75, \forall \beta, \forall x, \ \pi_\gamma^C(x) \geq \pi_\gamma^{CTP}(x) \geq \pi^{Prev}(x).$$

For any random sample larger than 5*, if we fix* γ *then* π_γ^C *is less specific than* π_γ^{CTP} *and* π_γ^{CTP} *is less specific than* π^{Prev}*.*

In this work, we focus on two-sided intervals because possibility distribution encoded two-sided intervals. The reviewed distributions can be used for different purposes in uncertainty management. Wallis [Wallis 51] used the Wald et al [Wald 46] normal tolerance limits to find tolerance intervals for linear regression. In the same way, we can use our γ-CTP distribution to build a probabilistic regression which encodes tolerance bounds of the response variable. Note that we are not restricted to possibilistic linear linear regression with homoscedastic and normal errors. We can also apply our γ-CTP and γ-DFCTP distributions for possibilistic non-parametric and parametric regression with homoscedastic or heteroscedastic errors.

3.5 Conclusion

We have, proposed different possibility distributions encoding different kinds of uncertainties. We also proposed a possibility distribution encoding confidence bands of the normal distribution which improves on the existing ones for all sample sizes. Building possibility distributions which encode tolerance intervals and prediction intervals are also new concepts that we introduced in this work. In future work, we propose to build, in the same way, the possibility distributions encoding distribution-free tolerance regions [Wald 43] and tolerance regions for the multivariate normal distribution [Krishnamoorthy 99]. The introduced γ-CTP distribution is used in [Ghasemi Hamed 12a] for possibilistic regression.

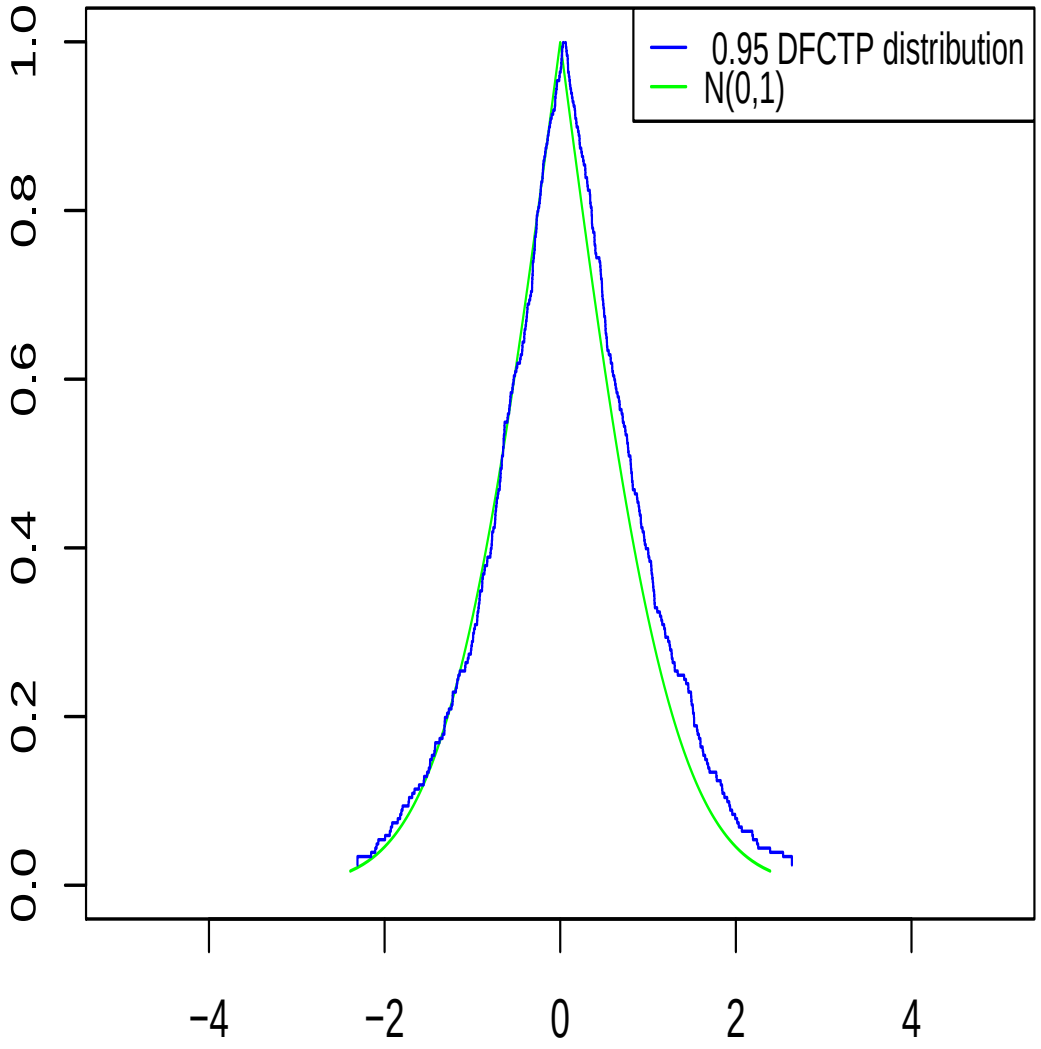

Figure 3.7: distribution-free 0.95-confidence tolerance possibility distribution for a sample set with size 450 drawn from $\mathcal{N}(0, 1)$.

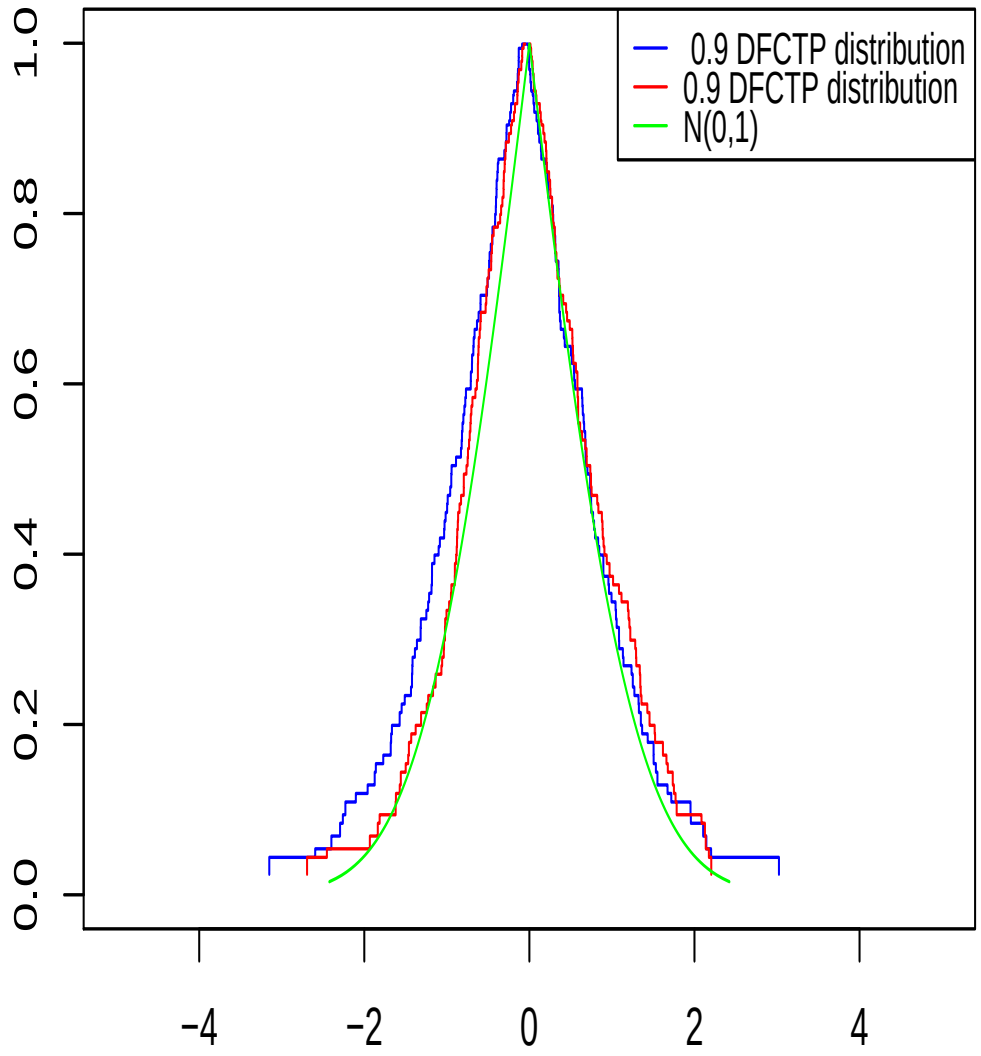

Figure 3.8: Two distribution-free 0.9-confidence tolerance possibility distributions for two sample sets of size 194 drawn from $\mathcal{N}(0, 1)$.

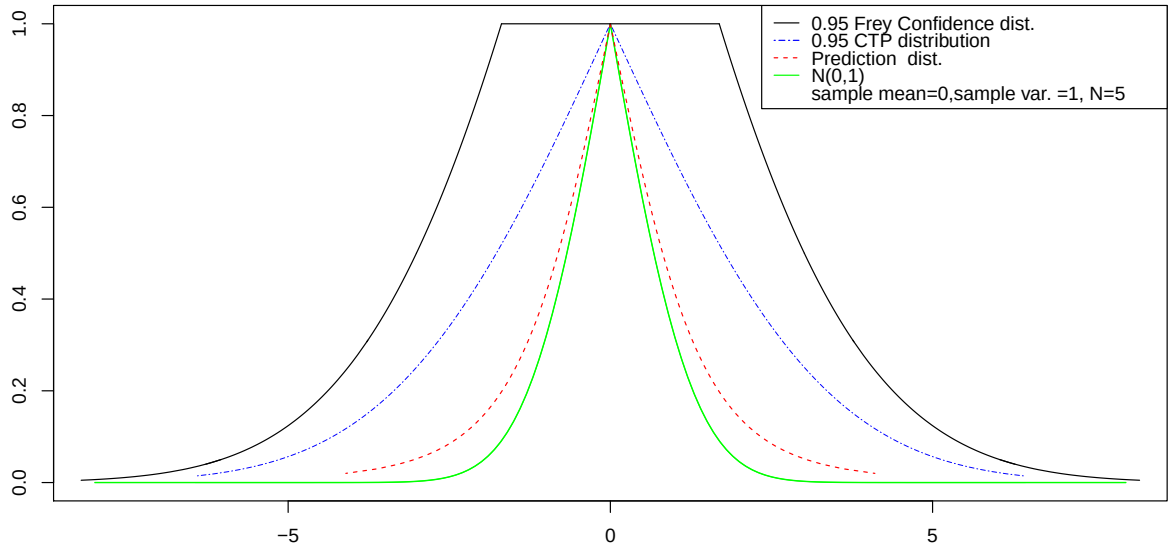

Figure 3.9: Comparing possibility distributions encoding Frey confidence band, tolerance intervals and prediction interval for a sample set with $n = 5$ drawn from a normal distribution having $(\bar{X}, S) = (0, 1)$.

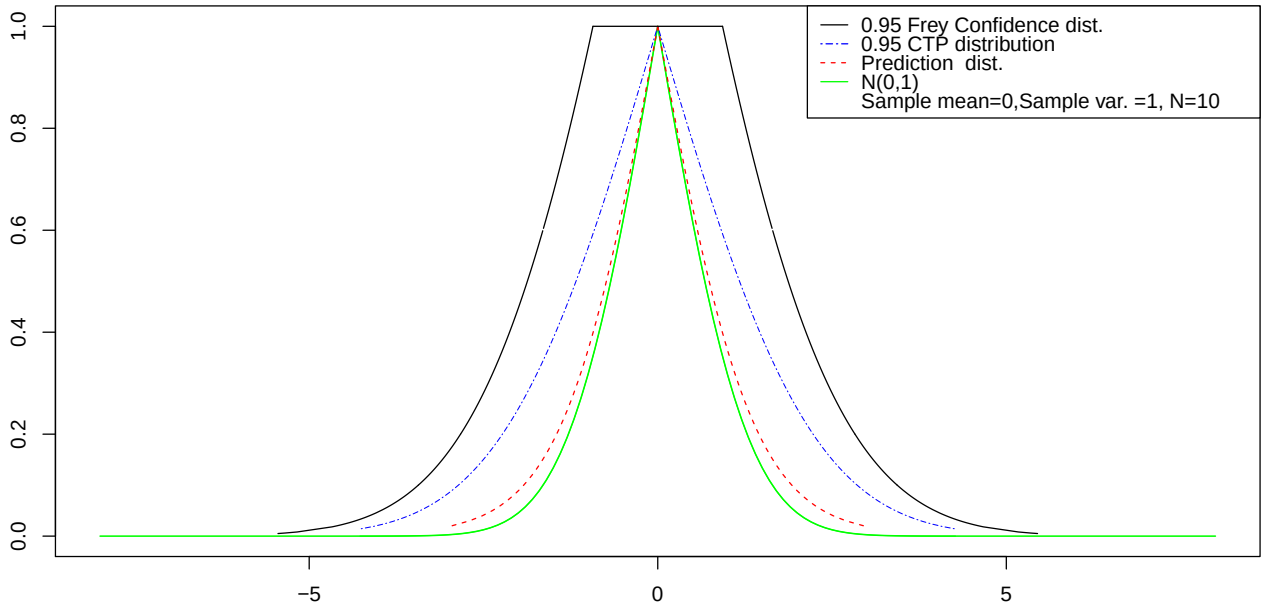

Figure 3.10: Comparing possibility distributions encoding Frey confidence band, tolerance intervals and prediction interval for a sample set with $n = 10$ drawn from a normal distribution having $(\bar{X}, S) = (0, 1)$.

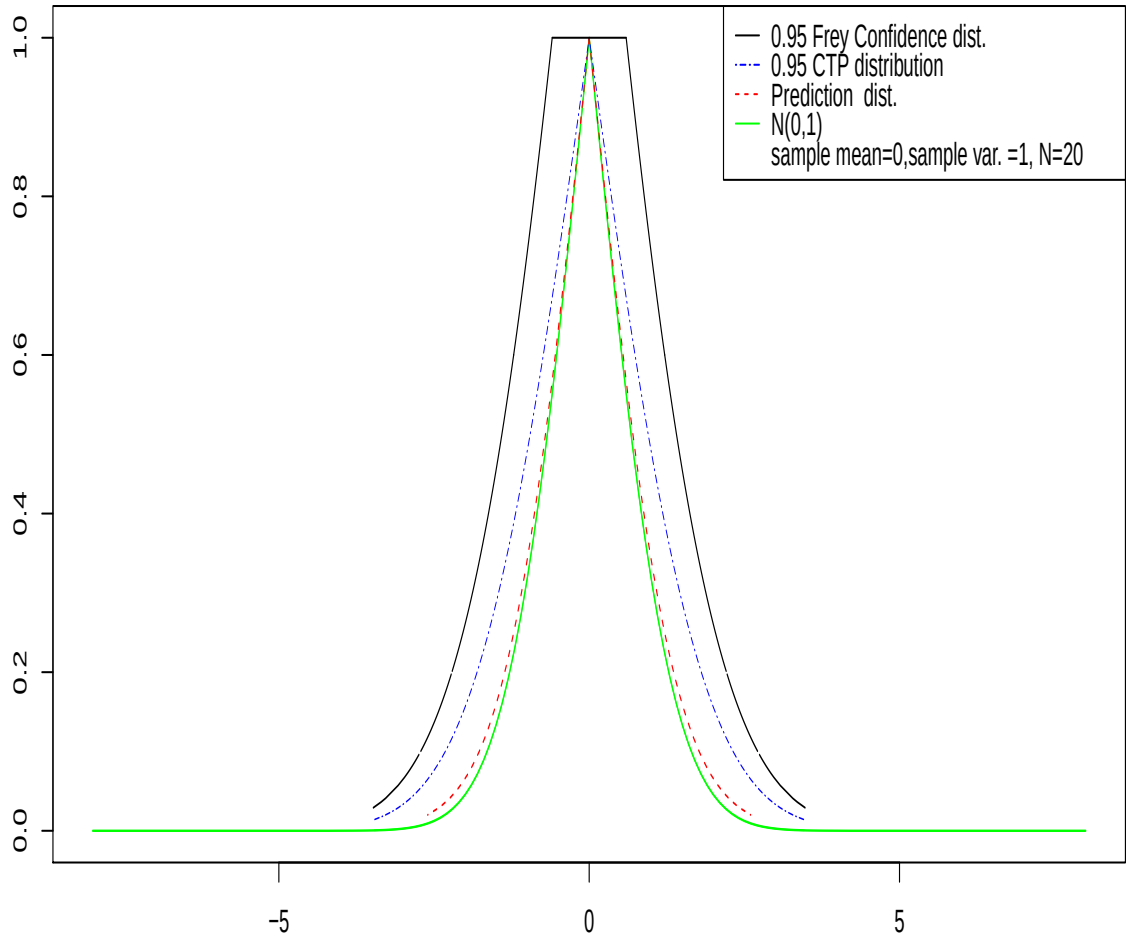

Figure 3.11: Comparing possibility distributions encoding Frey confidence band, tolerance intervals and prediction interval for a sample set with $n = 20$ drawn from a normal distribution having $(\bar{X}, S) = (0, 1)$.

Part II

Regression and Interval Prediction

Chapter 4

Regression

Contents

Regression analysis is a statistical technique for estimating the value of one variable as a function of independent variables. The estimated variable Y is called response variable or dependent variable Y and the independent variables x are also called predictors, explanatory variables or regressors. If there is one predictor, we have a simple regression and if each

predictor is a vector the problem is called multiple regression. Regression techniques are widely applied in science and engineering, they are used in problems like function estimation, financial forecasting, and time series prediction.

In the most general form a regression equation has three variables: the response variable Y , a deterministic function $f(x)$ and a random error ε, where $Y = f(x) + \varepsilon$. We divide statistical regression techniques into two categories. The first category estimates the mean of the random variable Y by $f(x)$ which as explained in the following section, is usually known as a least-squares model. The second one is called quantile regression. A function $f(x)$, based on predictor values, estimates conditional quantiles of Y. In each category, the regression function $f(\cdot)$ can be estimated with a parametric linear, a parametric non-linear or a non-parametric method. This results in linear, non-linear or non-parametric regression. We used the motorcycle dataset [Silverman 85] for illustrating these methods. This dataset is a well known non-linear regression dataset composed of 133 rows of accelerometer readings taken through time in an experiment to determine the efficacy of crash-helmets. This chapter does not contain any new contribution. It is a review on the regression models mentioned in the above paragraph. In this chapter, Y_i denotes the random response variable and y_i is an observation of the random variable Y_i.

4.1 Estimating the mean function

4.1.1 Regression

In fixed design regression, there are n pairs of observations $(x_1, Y_1), \cdots, (x_n, Y_n)$, where x_i is the vector of observations known as covariates and Y_i is the response variable. In other words, the random variable Y_i or $Y(x_i)$ follows a mean function $f(x_i)$ with a random error term ε_i defined as:

$$Y_i = f(x_i) + \varepsilon_i, \text{where } E(\varepsilon_i) = 0. \tag{4.1}$$

The model supposes that ε_i are mutually independent and identically distributed (iid) random variables. The goal is to estimate the mean function $f(\cdot)$ by $\hat{f}(\cdot)$, being as close as possible to the unknown function $f(\cdot)$. We could also treat the data as random where $(X_1, Y_1), \cdots, (X_n, Y_n)$ are random vectors. In this case, $f(x)$ is interpreted as the mean of Y conditional on $X = x$ as in (4.3).

$$E(Y|X = x) = f(x). \tag{4.2}$$

In this case, X_i are supposed iid and also independent from the ε_i's. This model is described by (4.3) where ε_i are iid and have zero mean and unit variance.

$$Y_i = f(X_i) + v^{\frac{1}{2}}(X_i)\varepsilon_i, \tag{4.3}$$

$$\text{where} E(\varepsilon_i) = 0, v(X_i) = Var(Y|X = x_i).$$

These models are different formulations of regression, however while working with local polynomial regression explained in (4.3.2) the formula remains the same for both and in this work we will refer to the fixed design approach. The usual assumption is to suppose that the variance of the error is the same everywhere. This is known as homoscedasticity and the opposite hypothesis (variable error variance) is known as heteroscedasticity.
In a least squares regression, the idea is to estimate the mean of $Y(x)$ by $\hat{f}(x)$ and based on some assumptions, described in 4.1.2, it results in finding the function that minimizes the Mean Squared Error (MSE), i.e. finding $\hat{f}(\cdot)$ that minimizes:

$$MSE(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

4.1.2 Mean Squared Errors (MSE) and Predictive Risk

Predictive risk is tightly coupled with the MSE. This notion is the most commonly used measure for tuning of hyper-parameters, model selection and inference. We will devote the current subsection to its definition. The risk of an estimator is the square of the difference between the true value of the parameter and its estimation. Given a fixed value of x, the mean of squared error for all values of the random variable $\hat{f}(x)$ is defined as the risk at point x:

$$MSE_{\hat{f}(x)} = RISK_{\hat{f}(x)} = E[(f(x) - \hat{f}(x))^2]. \tag{4.4}$$

It is well known that (4.4) can also be decomposed in bias and variance terms as in (4.5).

$$MSE_{\hat{f}(x)} = E[(\hat{f}(x) - f(x))^2] = Bias^2_{\hat{f}(x)} + V_{\hat{f}(x)} \tag{4.5}$$

$$\text{, where } Bias_{\hat{f}(x)} = E[\hat{f}(x) - f(x)] \tag{4.6}$$

$$\text{and } V_{\hat{f}(x)} = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2. \tag{4.7}$$

Average Mean Square Error of $\hat{f}(\cdot)$, or the average risk of $\hat{f}(\cdot)$, is the average of the mean squared error of $\hat{f}(\cdot)$ over all values of x, and it is used as an evaluation measure in regression problems.

$$\text{Average MSE } = \frac{1}{n}\sum_{i=1}^{n} E[(\hat{f}(x_i) - f(x_i))^2].$$

The average risk is related to the predictive risk. Let us first define the squared prediction error. The squared prediction error is the squared error of prediction for a new observation (x_i, Y^*) and it is defined as:

$$(Y^* - \hat{f}(x_i))^2 = (f(x_i) + \epsilon^* - \hat{f}(x_i))^2.$$

The predictive risk is:

$$\begin{aligned}\text{Predictive Risk} &= \frac{1}{n}\sum_{i=1}^{n} E[(Y^* - \hat{f}(x_i))^2] \\ &= \text{Average MSE } + \frac{1}{n}\sum_{i=1}^{n}\sigma^2(x_i),\end{aligned}$$

so we have:

$$\text{Predictive Risk} = \text{Average MSE } + c, \text{ where } c = \frac{1}{n}\sum_{i=1}^{n}\sigma^2(x_i); \tag{4.8}$$

$\sigma^2(x_i)$ is the variance of the response variable at x_i, and c is a constant. If the error variance $\sigma^2(x_i)$ is constant for all x_i, then

$$\text{Predictive Risk} = \text{Average MSE } + \sigma^2. \tag{4.9}$$

Hence based on (4.8), minimizing the predictive risk results in minimizing the average risk of the estimated regression function $\hat{f}(\cdot)$. *In a small to medium size dataset, leave-one-out or 10-fold cross validation MSE are well-known estimators of predictive risk.*

4.2 Linear Regression

Linear regression was the first type of regression analysis to be studied rigorously, and has been used extensively in practical applications. This is due to the fact that regression models which linearly depend on their parameters are easier to fit than non-linear regression models. In statistics, a linear model uses a linear function $f(x)$ to represent the relationship between a dependent random variable Y and a k-dimensional vector of predictor variables x. When we have a sample of n observations (x_i, y_i)[1], in most cases it is not possible to find a linear function $f(\cdot)$ of the k-dimensional input vector x for which $y_i = f(x_i)$ holds for all $i \in (1, \ldots, n)$. So this inequality is modeled through an error ε_i, which is an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Hence we have:

$$Y_i = f(x_i) + \varepsilon_i = x_i^T\beta + \varepsilon_i,$$

where x_i^T is the transpose of x_i, and β is a p-dimensional ($p = k + 1$) vector of parameters in the linear function $f(\cdot)$.
If we stack these n equations together and write them in vector form we have:

$$\mathbf{Y} = X\beta + \varepsilon, \tag{4.10}$$

[1] y_i is an observation of the random variable Y_i.

$$\mathbf{Y} = (Y_1, \cdots, Y_n)^T,$$
$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T,$$

where $\mathbf{Y}$ is the vector of response variables, $\mathbf{X}$ represents a matrix of all x_i, and ε is the vector of all errors. In this context, we look for the best estimate of y_i (minimizes the risk) written as $\hat{y}_i$. The response variable is estimated by the equation below in which $\hat{\beta}$ is an estimate of the true vector β:

$$\hat{y}_i = \hat{f}(x_i) = x_i^T \hat{\beta} \tag{4.11}$$

In the parameter estimation phase, we are searching for the vector of parameters $\hat{\beta}$ which fits a straight hyperplane through the set of n points in a way to minimize the sum of squared errors defined by (4.4).

4.2.1 Ordinary Least Squares problem (OLS)

The most common estimation method for a linear model is the OLS which is described in this section. The assumptions are stated below :

- The matrix $\mathbf{X}$ must have full column rank p, otherwise we have what is called perfect multicollinearity in the regressors. Methods for estimating parameters in linear models with multicollinearity have been developed, [Draper 79], [Tibshirani 96], [Efron 04], but they require additional assumptions.
- The regressors x_i are assumed to be error-free. It means that they do not have measurement errors. This otherwise leads to another problem known as errors-in-variables models.
- ε has the normal distribution:

$$\varepsilon \backsim \mathcal{N}(0, \sigma^2 I). \tag{4.12}$$

The last statement is one of the most common assumptions in practice and we did similarly. It also implies spherical errors:

- Homoscedasticity: $\forall i, Var[\varepsilon_i^2] = \sigma^2$. The inverse hypothesis, heteroscedasticity, is made when error terms do not have necessarily equal variance. Under such cases it might be better to use a weighted version of the OLS named Weighted Least Squares (WLS). WLS minimizes a weighted version of the sum of squared error terms, where each error term is weighted by a factor that indicates the precision of the information contained in the associated observation.
- Non-autocorrelation of errors $E[\varepsilon_i \varepsilon_j] = 0, \forall i \neq j$.

In this situation, the Gauss-Markov theorem states that minimizing the sum of squared residuals gives us the Best Linear Unbiased Estimator (BLUE). In other words, OLS fits a plane through the set of n vectors in such a way it that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted plane) as small as possible.

$$\hat{\beta} = \underset{\beta}{\text{Argmin}}(U^T U), \text{ where } U = y - X\beta$$

In OLS the BLUE estimator is found by:

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{4.13}$$

The vector $\hat{\beta}$ is distributed normally:

$$\hat{\beta} \backsim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}), \tag{4.14}$$

and $\hat{\sigma}^2$ which is the maximum likelihood estimator of σ^2 and the $\frac{n\hat{\sigma}^2}{\sigma^2}$ term has a chi-square distribution with $n - p$ degrees of freedom [Mendenhall 06].

$$\hat{\sigma}^2 = \frac{\hat{U}^T \hat{U}}{n}, \text{ where } \hat{U} = y - X\hat{\beta} \tag{4.15}$$

$$\frac{n\hat{\sigma}^2}{\sigma^2} \backsim \chi^2_{n-p}. \tag{4.16}$$

Figure 4.1 shows a simple linear model built on a dataset of 100 observations drawn from $\mathcal{N}(10 + 5x, 70^2)$. The red line represents the true mean function and the blue line is the OLS.

4.2.2 Weighted Least Squares (WLS)

The homoscedasticity assumption does not hold, even approximately, in every modeling application. In this such cases, OLS is not BLUE and it is better to use a weighted version of the OLS named Weighted Least Squares (WLS). WLS minimizes the wighted distance error term of each observation.

$$\arg\min_{\tilde{\boldsymbol{\beta}}} \sum_{i=1}^{n} w_i |y_i - f(x_i)|^2 = \arg\min_{\tilde{\boldsymbol{\beta}}} \left\| W^{1/2} (\mathbf{y} - X\tilde{\boldsymbol{\beta}}) \right\|^2,$$

where $w_i > 0$ is the weight of the i^{th} observation, and W is the diagonal matrix of such weights. The estimated parameter values are linear combinations of the observed values [Rao 99]:

$$\hat{\boldsymbol{\beta}} = (X^\top W X)^{-1} X^\top W \mathbf{y}. \tag{4.17}$$

It is shown [Rao 99] that in WLS, the estimator is BLUE if, when minimizing the weighted sum of squared residuals, we take each weight w_i equal to the reciprocal of the

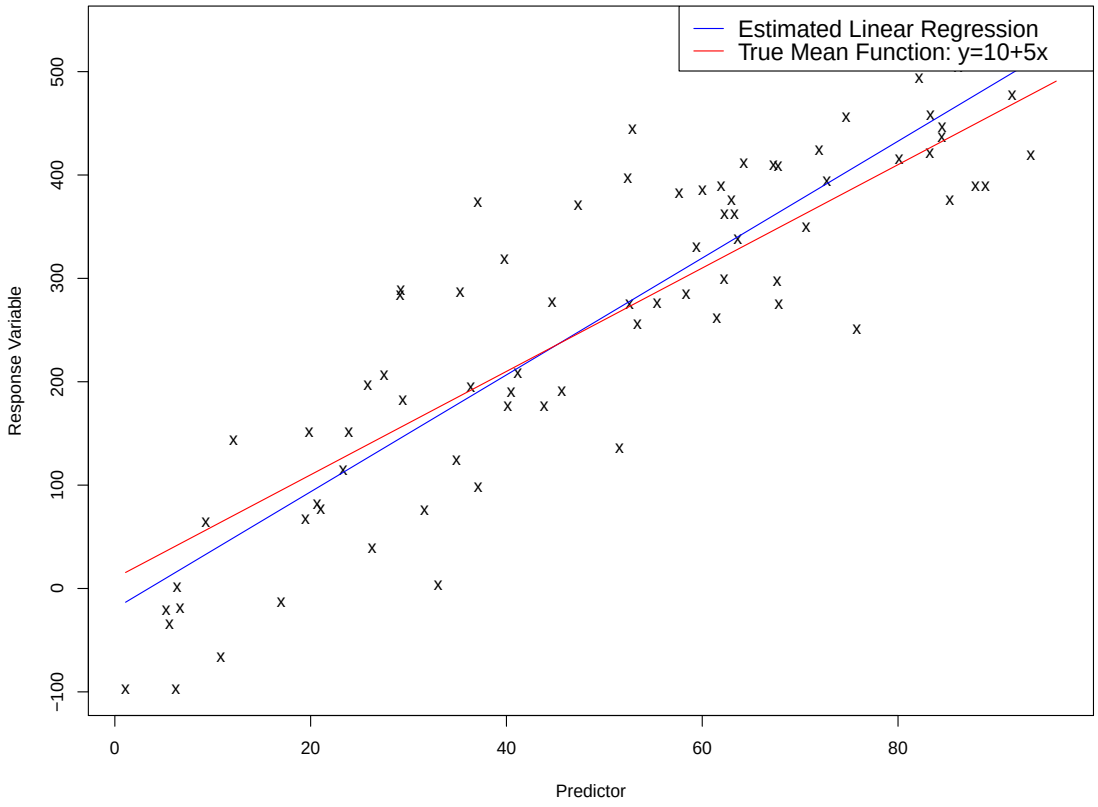

Figure 4.1: An OLSE model based on a sample set with $n = 100$.

variance of the measurement: $w_i = \sigma_i^{-2}$.
Each weight value indicates the precision of the information contained in the associated observation. Minimizing the weighted fitting criterion, allows the weights to determine the contribution of each observation to the final parameter estimates. Hence, the main advantage of WLS over other methods is its ability to handle regression cases having data points with varying quality. WLS assumes that the weights are known and we just want to optimize the parameters, which is almost never the case in real applications so we have to estimate them from the sample sets. Experience indicates that small differences between the estimated and true weights do not often affect a regression analysis or its interpretation. However, by using estimated weights from small numbers of replicated observations, the regression analysis result can be very badly and unpredictably affected [eng 11]. This can be the case when the weights for extreme values of the response variable are estimated using a few observations. Therefore it is important to use this method when weights can be precisely estimated relatively one to another [eng 11]. Chen and Shao [Chen 93] showed that at least three observations in each group of the same variance are required to obtain

WLS that asymptotically improve OLS. A more generalized form of WLS is the Generalized Least Squares (GLS), for more details see [Rao 99].

4.3 Local regression methods

4.3.1 State of the art

Non-parametric regression is a type of regression analysis in which the response value is not a predefined function of the predictor variables and vector of parameter θ which must be estimated from the data. As opposed to parametric regression, which is based on the construction on a model based on a training set, the prediction for a vector x is made by local estimation inside the training set. The motivation of non-parametric methods is their utility when dealing with complex models or when having non-linear and sometimes heteroscedastic data. Therefore, in such situations, exploiting the neighborhood of the input data to estimate the local distribution of response value may be justified. A number of monographs including Eubank (1988) [Eubank 99], Hastie and Tibshirani (1990) [Hastie 90], Hardle (1990) [Härdle 90], Wahba (1990) [Wahba 90] and Fan and Gijbels (1996) [Fan 96] have discussed this topic. Projection pursuit regression [Friedman 81], generalized additive models[Hastie 86], local polynomial regression [Cleveland 88], and Multivariate Adaptive Regression Spline (MARS) [Friedman 91] are common methods for nonparametric regression with multivariate predictor variables. These regression methods have been applied in multivariate case as well as in univariate case.

The idea of Local Polynomial Regression (LPR) first appeared in the statistical literature in Stone (1977) [Stone 77a] and Cleveland (1979) [Cleveland 79]. Cleveland (1979) [Cleveland 79], introduced Locally Weighted Regression (LWR) and a robust version of locally weighted regression known as Robust Locally Weighted regression Scatter plot Smoothing (LOWESS). LOWESS is an iterative version of LWR and the idea of LOWESS is to change the weight function defined in 4.19 so as to minimize outlier's impact. He states that this method is more convenient for regression datasets that have a non-normal error. Cleveland and Delvin (1988) [Cleveland 88] show that local polynomial regression can be very useful in real data modeling applications. They introduced "loess", which is a multivariate version of locally weighted regression. Their work includes the application of loess with multivariate predictor datasets. They introduce some statistical procedures analogous to those usually used in parametric regression. They also propose an ANOVA test for loess.

Fan (1992,1993) [Fan 92a, Fan 93] studied some theoretical aspects of local polynomial regression. Fan shows that Locally Weighted Linear Regression (LWLR) (or weighted local linear regression) is design adaptive. It adapts to random and fixed design as seen respectively in Equations (4.3) and (4.1). LWLR can be used in highly clustered as well as nearly uniform designz. He also shows the best local linear smoother has 100% efficiency among all possible linear smoothers, including kernel regression, orthogonal series and splines in minimax sense. Another important property of LWLR is its adaptation to

boundary points. As shown by Fan and Gijbels (1992) [Fan 92b], the LWLR estimator does not have boundary effects and therefore it does not require any modification at the boundary points. This is a very attractive property of these estimators, because in practice, a large proportion of the data can be included in the boundary regions. Then Ruppert and Wand (1994) [Ruppert 94] extended Fan's results on asymptotic bias and variance to the case of multivariate predictor variables. Hastie and Loader (1993) [Hastie 93] discussed the bias, boundary effect and derivative estimation in locally weighted regression.

4.3.2 Local Polynomial Regression (LPR)

Local Polynomial Regression (LPR) assumes that the unknown function $f(\cdot)$ can be locally approximated by a low degree polynomial. LPR fits a low degree polynomial model in the neighborhood (x_i) of the point x. The estimated vector of parameters used in the fitted LPR is the vector that minimizes a locally weighted sum of squares defined later in Equation (4.20). Once the local polynomial is fitted to x's neighborhood, $\hat{f}(x)$ (or $\hat{y}$), is estimated by evaluating the local polynomial with x as predictor variable value. Thus for each x a new polynomial is fitted to its neighborhood and the response value is estimated by evaluating the fitted local polynomial with the vector x as covariate. In general the polynomial degree (d) is 1 or 2; **for $d = 0$, LPR becomes a kernel regression and when $d = 1$ it changes to Local Linear Regression (LLR)**.

Definition of LPR

Suppose that the regression function $f(\cdot)$ at the point x can be approximated locally for x_i inside a neighborhood of x. The idea is to write the Taylor's expansion for x_i inside a neighborhood of x as follows [Fan 96]:

$$f(x_i) = \sum_{j=0}^{d} \frac{f^j(x)}{j!}(x_i - x)^j \equiv \sum_{j=0}^{d} \beta_j (x_i - x)^j. \tag{4.18}$$

Equation (4.18) models the regression function by a polynomial function. Thus, for every observation z in the neighborhood of x, we write (4.18) and estimate the vector $\beta = (\beta_0, \cdots, \beta_d)^T$ by the vector of parameters $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_d)^T$ which minimizes the locally weighted sum of squares defined in Equation (4.19), and where $\mathcal{K}_b(\cdot)$ represents a kernel function with bandwidth b. In fact, estimating $f(x)$ for the random design as well as for the fixed design results in the locally weighted polynomial regression expressed by the equation below [Fan 96]:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{(d+1)}}{\text{Argmin}} \sum_{i=1}^{n} \mathcal{K}_b(x_i - x)\left(Y_i - \sum_{j=0}^{d} \beta_j (x_i - x)^j\right)^2 \tag{4.19}$$

The above formula can be re-expressed as:

$$\sum_{i=1}^{n} w_i \Big(Y_i - \hat{f}(x_i) \Big)^2, \tag{4.20}$$

$$\text{where } w_i = \mathcal{K}_b(x_i - x) \text{ and } \hat{f}(x_i) = \sum_{j=0}^{d} \hat{\beta}_j (x_i - x)^j.$$

By re-writing (4.19) in vector notation, we obtain

$$(\mathbf{Y} - \mathbf{X_x}\beta)^T \mathbf{W_x} (\mathbf{Y} - \mathbf{X_x}\beta), \tag{4.21}$$

where $\mathbf{Y}$ is the vector of response variables and for each x, $\mathbf{X_x}$ and $\mathbf{W_x}$ are respectively its predictor matrix and weight matrix as in (4.22).

$$\mathbf{Y} = (Y_1, \cdots, Y_n)^T,$$

$$\mathbf{X}_{\mathbf{x}(n\times(d+1))} = \begin{pmatrix} 1 & (x_1 - x) & \cdots & (x_1 - x)^d \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_n - x) & \cdots & (x_n - x)^d \end{pmatrix}, \mathbf{W_x} = \text{diag}(\mathcal{K}(\frac{x_i - x}{b}))_{n\times n}. \tag{4.22}$$

The vector $\hat{\beta}_x$ minimizing this weighted sum of squares is provided by WLS:

$$\hat{\beta}_x = (\mathbf{X_x}^T \mathbf{W_x} \mathbf{X_x})^{-1} \mathbf{X_x}^T \mathbf{W_x} \mathbf{Y}, \tag{4.23}$$

and $\hat{f}(x)$ becomes a linear smoother as in (4.24).

$$\hat{f}(x) = \sum_{i=1}^{n} a_i(x) Y_i, \tag{4.24}$$

$$\text{where } a(x) = I_1^T \hat{\beta}_x \text{ and } I_1^T = (1, 0, \cdots, 0).$$

We can also write the fitted values in vector notation as in Equation (4.25), where L is the projection matrix or the smoother matrix in which its $L_{ij} = a_j(x_i)$, and $\hat{\mathbf{f}}$ is the vector of fitted values.

$$\hat{\mathbf{f}} = L\mathbf{Y}, \tag{4.25}$$

$$\hat{\mathbf{f}} = (\hat{f}(x_1), \cdots, \hat{f}(x_n))^T.$$

Note that (4.23) works for single variate regression. When it comes to multivariate LPR with p predictor variables, the final d columns of $\mathbf{X_x}$ are repeated for each covariate. Hence, $\mathbf{X_x}$ becomes a $n \times (p \times d + 1)$ matrix, $\hat{\beta}_x$ a vector of $(p \times d) + 1$ element and the kernel function a multivariate kernel.

Kernel function:

In kernel regression or in LPR, a kernel function $\mathcal{K}(\cdot)$ is used to weight the observations. It is chosen so that observations closer to the fitting point x have bigger weights and those far from x have smaller weights. If $\mathcal{K}(\cdot)$ is a kernel, then $\mathcal{K}_b(\cdot)$ is also a kernel function.

$$\mathcal{K}_b(u) = \frac{1}{b}\mathcal{K}(\frac{u}{b}), \text{ where } b > 0.$$

Here, b, known as the bandwidth, is a constant scalar value used to select an appropriate scale for the data. A kernel function is a non-negative real-valued integrable function $\mathcal{K}(\cdot)$ with the properties listed below [Cleveland 79]. Almost all kernel function respect the first three properties, so they become probability density functions. The last property limits the neighborhood and this helps to achieve better computing performance. For more explanation about the weight function properties see [Cleveland 79].

$$\forall u, \mathcal{K}(-u) = \mathcal{K}(u)$$
$$\forall u, \mathcal{K}(u) \geq 0$$
$$\int_{-\infty}^{+\infty} \mathcal{K}(u)\, du = 1,$$
$$\mathcal{K}(u) > 0, |u| < 1$$

In the following, you can see some of the most common kernel choices [Li 07]. Note that $I(\cdot)$ is the indicator function.

- Gaussian: $\mathcal{K}(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$,
- Tricube: $\mathcal{K}(u) = \frac{70}{81}(1 - |u|^3)^3\, \mathbf{I}_{\{|u| \leq 1\}}$,
- Epanechnikov: $\mathcal{K}(u) = \frac{3}{4}(1 - u^2)\, \mathbf{I}_{\{|u| \leq 1\}}$,
- Uniform: $\mathcal{K}(u) = \frac{1}{2}\, \mathbf{I}_{\{|u| \leq 1\}}$,
- Triangle: $\mathcal{K}(u) = (1 - |u|)\, \mathbf{I}_{\{|u| \leq 1\}}$,
- Quartic (biweight or bisquare): $\mathcal{K}(u) = \frac{15}{16}(1 - u^2)^2\, \mathbf{I}_{\{|u| \leq 1\}}$,
- Triweight: $\mathcal{K}(u) = \frac{35}{32}(1 - u^2)^3\, \mathbf{I}_{\{|u| \leq 1\}}$,

- Cosine: $\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)\mathbf{1}_{\{|u|\leq 1\}}$

For multivariate LPR, the kernel function $\mathcal{K}_B(\cdot)$ is a function of p variables. In this case B is a symmetric positive definite $p \times p$ matrix and $|B|$ denotes its determinant. It can be redefined as:

$$\mathcal{K}_B(u) = \frac{1}{|B|}\mathcal{K}(B^{-1}u). \tag{4.26}$$

In practice, one can normalize or standardize all the predictors and then use the following kernel:

$$\mathcal{K}_b(u) = \frac{1}{b}\mathcal{K}\left(\frac{D(u)}{b}\right), \tag{4.27}$$

where $D(\cdot)$ is a distance function like the L_2-norm. Some authors including [Cleveland 79] and [Cleveland 88], took the K-nearest neighbors of x as its neighborhood. In this case, for each x, $b = D_k(x)$, where $D_k(x)$ is the distance from the K-th nearest neighbors (the farthest neighbor) from the point x. For a detailed discussion see [Atkeson 97].

Bandwidth Selection

A popular bandwidth selection method is the Leave-One-Out (LOO) technique suggested in Stone (1977) [Stone 77a] which chooses the following bandwidth b:

$$b = \text{Argmin}\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(x_i))^2, \tag{4.28}$$

where $\hat{f}^{-i}(x_i)$ is the estimation without using the i^{th} observation obtained by Equation (4.24). Estimating the bandwidth by LOO is a time-consuming task, so it is common to minimize the K-fold cross validation score with $K = 5$ or $K = 10$; this leads to an approximation of LOO. Plug-in bandwidth is another smoothing strategy which is a formula for the asymptotically optimal bandwidth. The plug-in bandwidth requires several unknown quantities that must be estimated from the data. In section 4.2 of Fan and Gijbels (1996) [Fan 96], a plug-in bandwidth for linear weighted local linear regression is defined. One of the required parameters for this estimator is $f(\cdot)$'s second derivative which is more difficult to estimate than $f(\cdot)$ itself. In this work we use 10-fold cross validation to find the best bandwidth of our dataset.

Computational Complexity

In this part, we consider the computational complexity of LPR. We consider that the dataset is sorted and we ignore this computational complexity. In the most naïve case, LPR takes $O(n^2)$ operations. This is because the computation of weights at each point is $O(n)$. If we use kernel bounded support the computation takes $O(fn^2)$, where f is the fraction of

the whole dataset inside the neighborhood [Cleveland 79]. It results in $f = K/n$ for models using K-nearest neighbors as their neighborhood selection method like in [Cleveland 79] and [Cleveland 88]. Binning and updating algorithms are two categories of fast computation algorithms. A comparison of these fast implementations is made in [Fan 94]. They made a comparison between the two fast implementations and the naive version. The comparison is performed carefully with various settings, machine and softwares. The observed speed improvement factor is above hundreds for large sample size, for both methods. However neither the binning nor the updating algorithm dominates. The updating method has some problems of stability. The difference between the binned version and naïve implementation is small, and negligible from a practical point of view.
The binning method is an approximation of the direct method which reduces the number of kernel evaluations, based on the fact that many of these evaluations are approximately the same. The idea is to create an equally spaced grid of the dataset, so each grid point represents a bin. The dataset is modified by assigning each pair (x_i, Y_i) its nearest grid point $x_j(i)$. The modified dataset is summarized as:

$$\{(x_j, \overline{Y}_j, c_j), j = 1, \cdots, g\},$$

where $\overline{Y}_j$ denotes the bin average and c_j the bin counts.

$$\overline{Y}_j = \text{average}(Y_i, Y_i \in bin_j),\ c_j = \text{number of instances in } bin_j.$$

Note that the only approximation is the process of replacing each x_i by its nearest grid point. Then the estimation is done using the modified dataset. Thus the computational complexity of building a binning-based model is $O(n)$, and the evaluation is just done on the grid points which can reduce the complexity to $O(g)$, where g is the number of grid points. But if the model is built with grid points and then each true value of x_i is evaluated, rather than its nearest grid point, the evaluation complexity will be $O(ng)$.

The updating method relies on computing the average recursively. Each average is computed by updating the previous average. This concept is exploited in different ways by Friedman (1984) [Friedman 84] and Cleveland (1979) [Cleveland 79]. Gasser and Kneip (1989) [Gasser 89] introduce the updating idea for polynomial kernels, then this concept is developed more in [Fan 94]. The updating procedure principally refers to expanding the polynomials into expressions which can be calculated quickly by recursive updating. The updating version of LPR is of $O(n)$ complexity. The algorithm may suffer from numerical instability due to rounding errors. See [Seifert 94] for a solution to this numerical instability. For more details, see [Fan 94].

4.3.3 K-Nearest Neighbors (KNN)

K-nearest Neighbors (KNN) is a local regression method. For each query x, it takes the weighted average of response values of the neighborhood around x as an approximation to $f(x)$. The neighborhood of x are points in the predictor space which are nearest to x

than others, and the size of this neighborhood is controlled by the bandwidth which is the number K. KNN is a version of Local Polynomial Regression, described by Equation (4.20), where the polynomial degree is zero $d = 0$ and its weights are calculated with Equation (4.27) where $D(\cdot)$ is the Euclidean distance in the predictors space and b is the distance between the input vector x and its K^{th} nearest neighbor. The KNN estimator is a kernel smoother and can be also defined as:

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} \mathcal{K}_b(D(x, x_i))Y(x_i)}{\sum_{i=1}^{n} \mathcal{K}_b(D(x, x_i))}. \tag{4.29}$$

In fact, KNN is a specialized form of the Nadaraya-Watson [Nadaraya 64, Watson 64] kernel estimator in which the bandwidth b is not constant and depends on the distance between input vector x and its K^{th} nearest neighbor. Usually, the size of the neighborhood, K, has to be fixed before the learning phase and it will be constant for all the input vectors.

4.3.4 Loess

Loess was introduced by Cleveland and Delvin [Cleveland 88], and is a multivariate version of LOWESS [Cleveland 79], which is another version of LPR. Loess is described by (4.20), where the polynomial degree is one $d = 1$ or two $d = 2$. For the bandwidth selection and weight calculation, loess is similar to KNN. Its weights are calculated with (4.27) where, $u = (x_i - x)$, $D(\cdot)$ is u's L_2-norm in the predictor space and b is the Euclidean distance between the input vector x and its K^{th} nearest neighbor. The weight function chosen by Cleveland and Delvin [Cleveland 88] was the Tricube kernel, however any other weight function that satisfies the properties listed in the kernel definition could also be used.

In this work, we use linear loess as our non-parametric smoother function. Therefore, for each input vector x, we infer the vector of parameter $\hat{\beta}_x$ from the training set as in (4.23), where $d = 1$. As we can see in (4.30), for each prediction the locally weighted linear regression problem is converted to a WLS in which the weights are estimated by a kernel function.

$$\hat{\beta}_x = \arg\min \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2, \tag{4.30}$$

where $w_i \geq 0$ is the weight of the i^{th} observation. Figure 4.2 uses the motorcycle dataset from [Silverman 85] to compare KNN with linear loess. We can see that linear Loess gives a slightly smoother function.

4.4 Quantile Regression (QR)

Koenker and Bassett (1978) [Koenker 78], introduced quantile regression in which we find estimation of conditional quantiles of the response variable Y given $X = x$. Least squares regression estimates the conditional *mean* of the response variable based on given values of the independent variables, whereas quantile regression extends the regression model to the

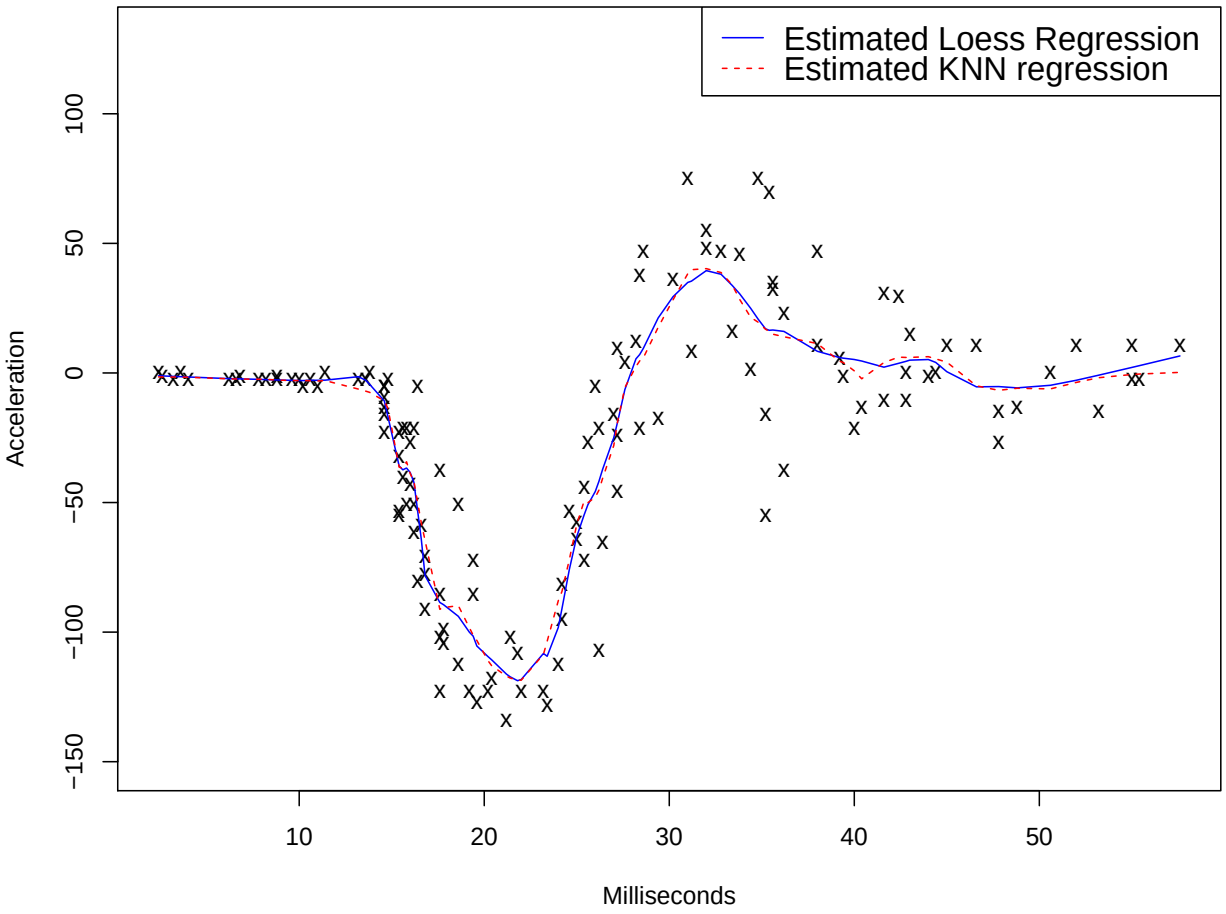

Figure 4.2: Comparing Loess regression with $k = 20$ and KNN regression with $k = 12$ for the motorcycle data from [Silverman 86].

conditional quantiles of the prediction variable (given the predictor values). We focus to find 50, 75 or 95 percentile of the conditional distribution of Y, given $X = x$ ($F(Y|X = x)$).

Least-squares methods are used much more than quantile regression and Koenker [Koenker 05] mentioned three possible reasons for this dominance:

- The linear estimators in least-squares methods are computationally tractable.
- If the noise is Gaussian, in some cases the least-squares method becomes the unbiased minimum variance estimator.

- Recent observations show that least-squares methods are a general approach for estimating the mean function.

However, the mean or its estimation is not sufficient. Even if we are interested in having a statistical analysis on single sample, we need to obtain other measures like skewness, kurtosis, density estimation. This is why quantile regression techniques can help us to have a clearer picture of the regression problem. It is also important to note that quantile regression is much more flexible than least squares regression when dealing with heterogeneous conditional distributions, because it makes no distributional assumption about the error term in the model and just provides a conditional distribution of the prediction given the predictor values [Koenker 01].

4.4.1 Linear Quantile Regression (LQR)

Linear quantile regression considers the problem of estimating a vector of unknown parameters, β, from an identically distributed random sample $(Y_1, \cdots, Y_n)$ where each random variable comes from

$$Y_i = x_i^T \beta + \varepsilon_i, \text{ where } F_{\varepsilon_i}^{-1}(\tau) = 0, \tag{4.31}$$

and ε_i comes from an unknown distribution $F_\varepsilon(\cdot)$. Quantile regression does not assume that the shape of F_ε is known. In the case of a Gaussian F_ε, Rao demonstrated that the least squares estimator is its minimum variance unbiased estimator. However when the distribution is not normal, least-squares linear regression can in many cases obtain very poor results [Koenker 78]. In fact, linear quantile regression is also an alternative robust estimator for linear models.

The τth conditional quantile distribution of Y is a linear function where:

$$Q_\tau(y|x) = x_i^T \beta + F_\varepsilon^{-1}(\tau),$$

which is written:

$$Q_\tau(y|x) = x_i^T \beta_\tau. \tag{4.32}$$

Having the observations $(x_i, y_i)\,(i = 1, \cdots, n)$, we can estimate β_τ by $\hat{\beta}_\tau$ by solving the following optimization problem:

$$\beta_\tau = \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T \beta), \tag{4.33}$$

where

$$\rho_\tau = r(\tau - I(\tau < 0)), \tag{4.34}$$

and $I(\cdot)$ is the indicator function. Equation (4.33) can be reformulated to a linear programming problem where $\mathbf{X}$ is the $n \times p$ matrix of predictors and $\mathbf{Y}$ denotes the vector of response values [Koenker 05]:

$$\min(\tau 1_n^T U + (1 - \tau) 1_n^T V), \tag{4.35}$$

with constraints:

$$\begin{gathered}\beta \in \mathbb{R}^p, \\ (U, V) \in R_+^{2n} \text{ and} \\ \mathbf{X}\beta + U - V = \mathbf{Y}.\end{gathered}$$

If the errors on observations are iid, Koenker and Basset [Koenker 78] have shown that $\hat{\beta}_\tau$ is asymptotically normal. If ε_i are independent but not identically distributed then the asymptotic covariance is:

$$V_\tau = (\tau(1-\tau))(\mathbf{X}^T F \mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T F \mathbf{X})^{-1}, \tag{4.36}$$

where $F = \text{diag}\{f_1(0), \cdots, f_n(0)\}$, and $f_i(\cdot)$ is ε_i's probability density function; in a model with iid errors the much more variance becomes [He 96]:

$$V_\tau = \frac{\tau(1-\tau)}{f^2(0)}(\mathbf{X}^T F \mathbf{X})^{-1}. \tag{4.37}$$

Note that if we want two or more quantiles from a finite dataset, the estimated quantile regressions may cross or overlap with each other, which is called as *quantile crossing*. This phenomenon occurs because each of the quantile functions has been estimated independently [Koenker 05]. Quantile crossing is a troublesome problem, because it is in contradiction with the semantic of the quantile estimation problem. Fortunately, this problem can be avoided by estimating all of the selected quantile functions the same time enforcing the *non-crossing* constraint. ***However after enforcing this constraint the conditional quantile estimator may not converge to the true conditional quantile.***

4.4.2 Non-linear Quantile Regression

By surveying the non-linear quantile regression literature, we can observe that it has received much less attention than the linear quantile regression. Konenker and Park [Koenker 96] have developed an interior point algorithm for non-linear quantile regression. This method is implemented by the *nlrq* function in the *R*'s *quantreg* package. The problem consists of minimizing

$$\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - g(x_i, \theta)), \tag{4.38}$$

where $g(\cdot, \theta)$ is considered to be continuously differentiable in θ. For more details see [Koenker 96, Koenker 05]. Takeuchi et al. [Takeuchi 06] have also proposed a kernel-based quantile regression method. They compared their method on several benchmarks and artificial datasets to linear quantile regression and the spline quantile regression, introduced by Koenker et al. [Koenker 94b] (it is provided by the *rqss* function in *R*'s *quantreg* package). Their method consists of solving a quadratic programming problem. Their experiments show the feasibility of their kernel-based quantile regression and compared

it with linear and spline based quantile regression. The reader can find this method implemented in the *R*'s *kernlab* package. Figure 4.3 shows Takeuchi et al.'s method [Takeuchi 06] applied to the motorcycle dataset. In this figure the training set and the validation set are the same. We can see that the proposed kernel based quantile regression fits well the nonlinear structure of the data but it suffers from the quantile crossing problem. Figure 4.4 is also an application of Takeuchi et al.'s method to the motorcycle dataset. But in this case the method is applied in a 10-fold cross validation schema. We can easily see that in this case, the kernel method becomes much less reliable and its crossing quantile effects get much more stressed than in the previous example.

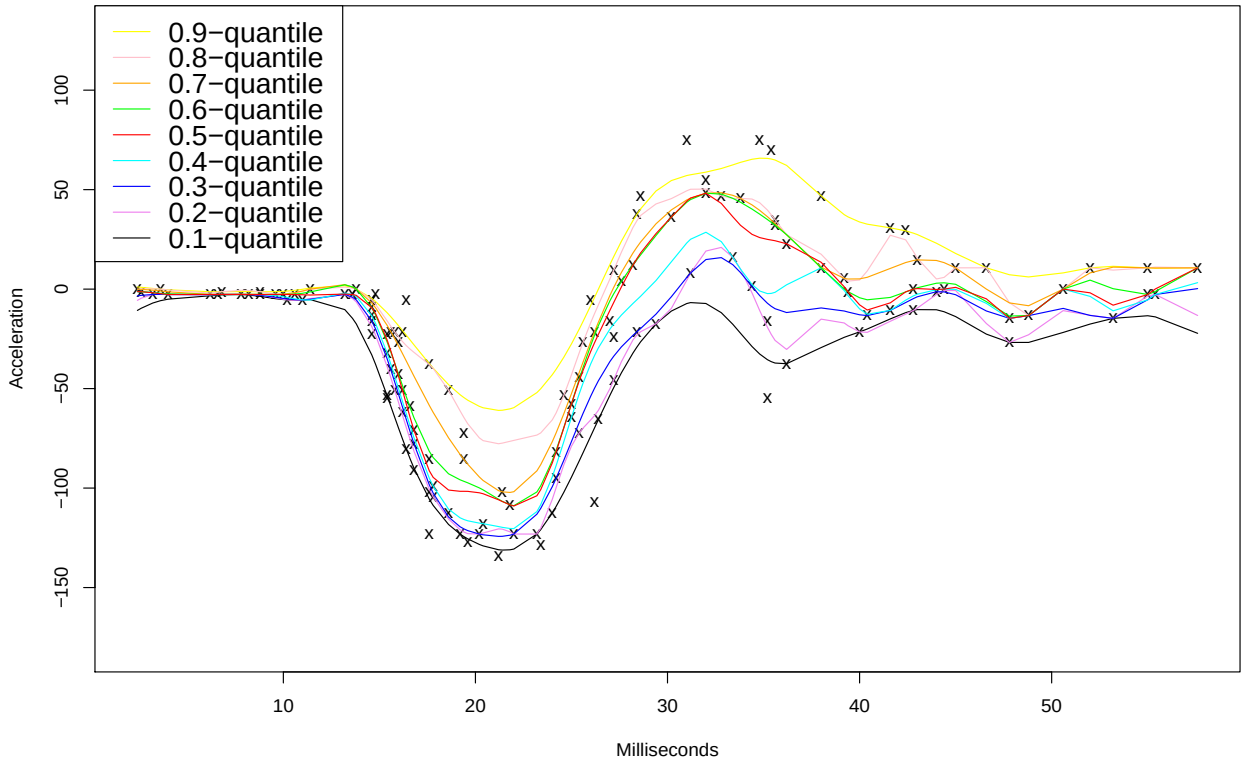

Figure 4.3: Kernel-based non-linear qunatile regression applied to the motorcycle dataset [Silverman 86].

4.4.3 Non-parametric Quantile Regression

There is an extensive literature on non-parametric quantile regression. Stone [Stone 77b] considers the K-nearest neighbors quantile regression. He establishes its consistency and

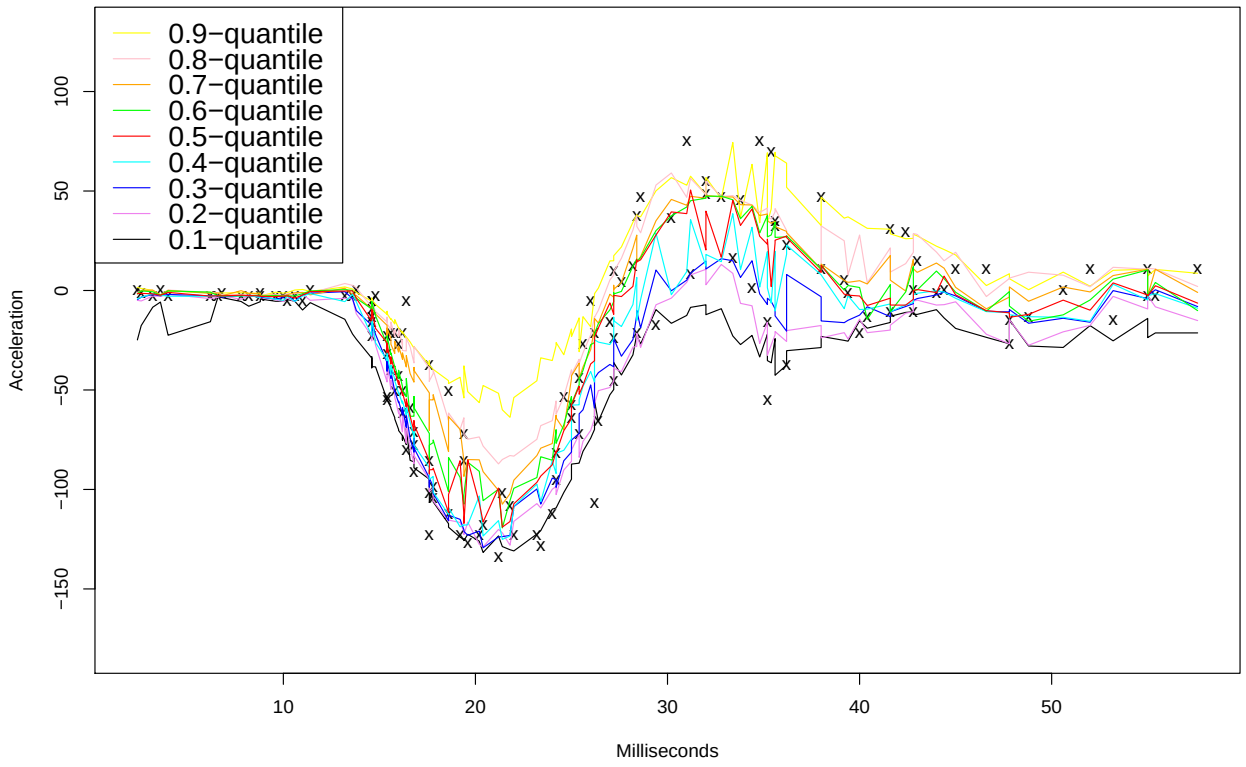

Figure 4.4: Kernel-based non-linear qunatile regression applied in a 10-fold cross validation schema to the motorcycle dataset [Silverman 86].

rate of convergence. Bhattacharya et Gangopadhyay [Bhattacharya 90] studied kernel and nearest-neighbor quantile estimators. In order to consider the asymptotics of these estimates, they obtained Bahadur type [Bahadur 66] representations of the sample conditional quantiles. Then Chaudhuri [Chaudhuri 91] introduced locally polynomial quantile regression. As we saw in Section 4.3.1, kernel and nearest-neighbor estimators are a version of local polynomial estimator having a polynomial of degree zero. The Chaudhuri [Chaudhuri 91] conditional quantile estimator differs from the the Fan and Gijbels [Fan 92b] conditional mean estimator, by its loss function. The mean estimator uses the squared errors as its loss function and the quantile estimator uses the Pinball loss function described previously in Equation (4.34). Equation (4.39) describes a conditional local polynomial quantile estimator for a one dimensional fixed design quantile regression model. If the predictors are in the p-dimensional

space, then the parameter β must be optimized in $\mathbb{R}^{p\times(d+1)}$.

$$\min_{\beta\in\mathbb{R}^{d+1}} \sum_{i=1}^{n} w_i \rho_\tau \Big(Y_i - \sum_{j=0}^{d} \hat{\beta}_j (x_i - x)^j \Big), \tag{4.39}$$

$$\text{where } w_i = \frac{1}{b} \mathcal{K}(\frac{x_i - x}{b}).$$

Figure 4.5 shows the local linear quantile regression applied to the motorcycle dataset. In this figure the training set and the validation set are the same. We can see that the proposed kernel based quantile regression fits quite well the nonlinear structure of the data but it suffers much less from the quantile crossing problem than the non-linear model in figure 4.3. Figure 4.6 is also an application of Chaudhuri's method to the motorcycle dataset. But in this case the method is applied in a 10-fold cross validation schema. We can easily see that in this case, the kernel method becomes much less reliable, its crossing quantile effects appears more and the quantile estimation function becomes much less smooth.

4.5 Other interval regression methods

4.5.1 Methods with an optimization point of view

Interval regression has been studied based on several contexts. Tanaka [Tanaka 87] was the first to propose possibilistic regression, which is reminiscent of quantile regression. The goal of this approach is to associate the data with a pair of upper and lower regression functions, while minimizing the total spread of the output coverage. Then Ishibuchi and Tanaka [Ishibuchi 90] proposed several reformulations of the linear interval regression model with interval data. This work reformulates the problem as a linear programming problem. In another work [Ishibuchi 92], they used neural networks to handle nonlinear interval regression models with interval data. Their work consists of employing two back-propagation networks (BPNs); one network identifies the upper side of the interval valued response variable and the other one finds its lower side. Ishibuchi et al. [Ishibuchi 93] use one interval neural network to represent both the upper and lower sides of the interval response variable. This work first proposes an architecture of neural networks with interval weights and interval biases. This neural network maps an input vector of real numbers to an output interval. Cheng and Lee [Cheng 01] proposed to use radial basis function network in fuzzy regression analysis without predefining any functional relationship between the covariates and the response variable. The proposed approach is a fuzzification of the connection weights between the hidden and the output layers. This fuzzy network is trained by a hybrid learning algorithm, where c-mean clustering method and the K-nearest-neighbor heuristics is used for training the parameters of the hidden units and linear possibilistic programming is used for updating the weights between the hidden and the output layers. Huang et al. [Huang 98] introduced robust interval regression for neural networks. They proposed two robust learning algorithms for determining a robust nonlinear interval regression model. The two robust algorithms are derived in a similar manner to the back-propagation (BP)

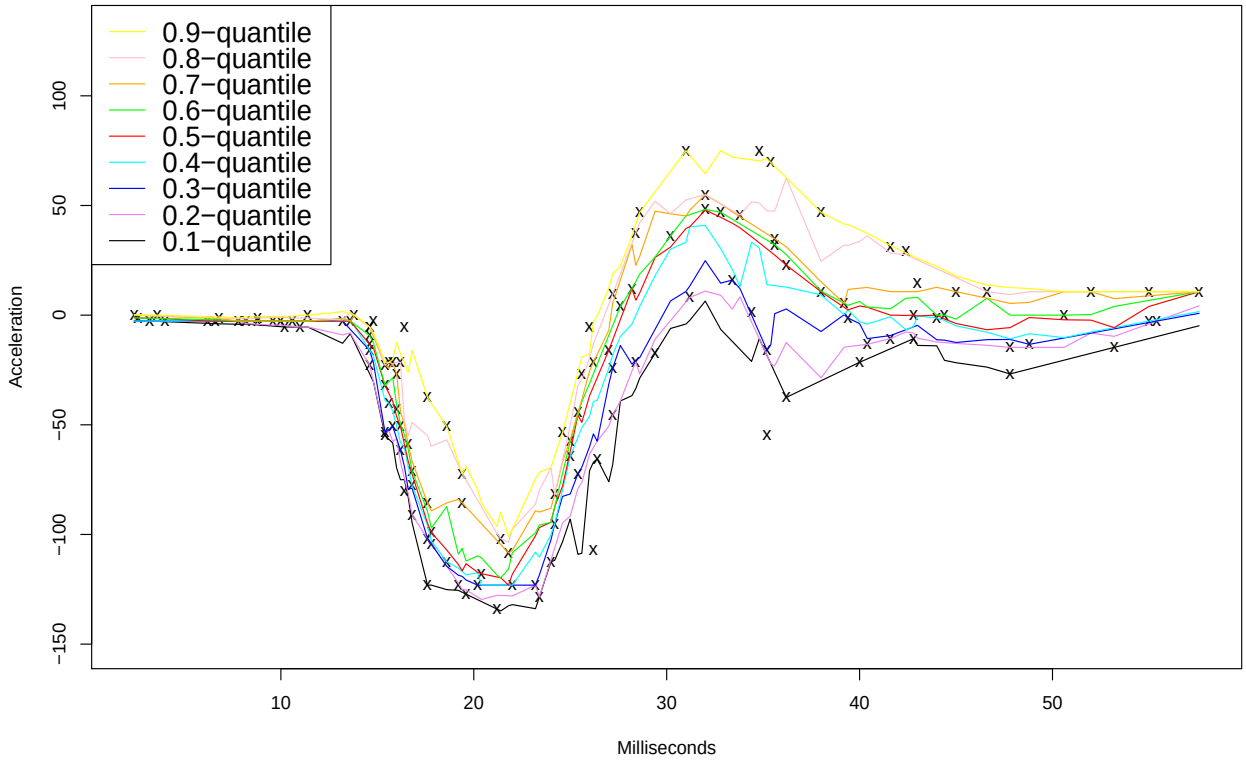

Figure 4.5: Local linear quantile regression with a bandwith of 20-nearest neighbors applied to the motorcycle dataset [Silverman 86].

algorithm. Peters [Peters 94] proposed a fuzzy LP based method for constructing a robust fuzzy linear regression model. Interval regression with Support Vector Machines (SVM) has been introduced by [Jeng 03] and [Hong 03]. Jeng et al. [Jeng 03] proposed Support Vector Interval Regression Network (SVIRN) for interval regression analysis. SVIRN uses a pair of radial basis function networks. One network identifies the upper side of interval valued response variable, and the other network finds its lower side. In the proposed method, the SVIRN approach with the ϵ-insensitive loss function is used to obtain the initial structure of SVIRNs. Then, a BP learning algorithm is employed to adjust the two networks. SVIRN is a robust interval regression method for interval numeric and interval output data. Hong and Hwang [Hong 05] proposed interval regression with support vector machine using a quadratic loss function.

Petit-Renaud and Denoeux [Petit-Renaud 04] were the first to propose a regression analysis approach for imprecise and uncertain data. They called their model "EVREG" (EVidential REGression) model which is a version of the KNN algorithm that uses a

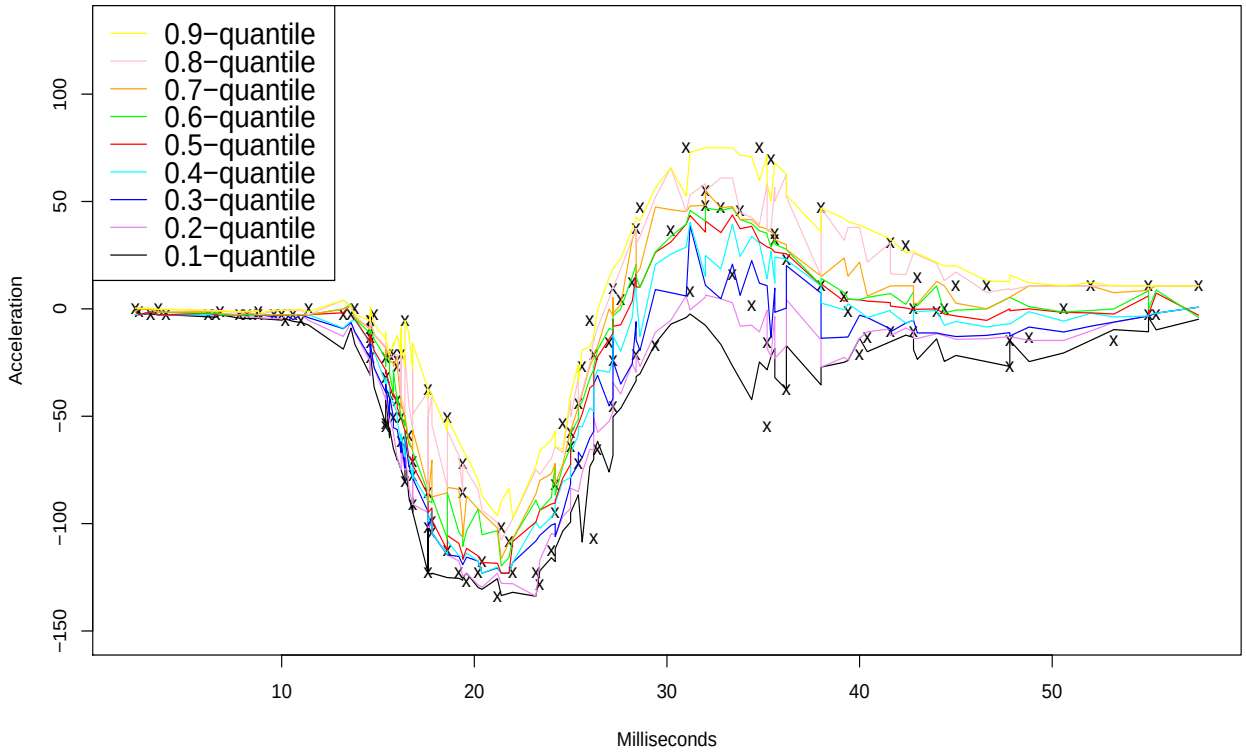

Figure 4.6: Local linear quantile regression with a bandwith of 20-nearest neighbors applied in a 10-fold cross validation schema to the motorcycle dataset [Silverman 86].

fuzzy belief function to represent the input and output data. Zhi-gang and Wang [Su 13] investigate the multiple linear and nonlinear regression of belief function on interval-valued variables. They also extended the EVREG model to deal with belief function on interval-valued variables. Their main contribution is the proposed parametric and nonparametric evidential regression models that deal with regression of imprecise and uncertain data, represented as a belief function with finite interval-valued focal elements.

In the context of possibility theory, Serrurier and Prade [Serrurier 07] considered imprecise regression for triangular and trapezoidal fuzzy sets. They use the simulated annealing algorithm to find a model that has the best tradeoff between faithfulness with respect to data and (meaningful) precision. In the same context, we proposed "Possibilistic KNN regression using tolerance intervals" [Ghasemi Hamed 12a]. This is the first imprecise regression approach with crisp-input and crisp-output data which address the reliability of an imprecise model by its coverage of the response values. The idea is to use tolerance intervals to build the maximal specific possibility distribution that bounds population quantiles of

the unknown conditional distribution of the response value. The interval regression method that is introduced is based on KNN and takes advantage of our possibility distribution in order to choose, for each instance, the value of K which will be a good trade-off between precision and uncertainty due to the limited sample size. This work is explained in section 7. Celmins [Celmins 87] and [Diamond 88] have another point of view for fuzzy regression, which is in contrast to all the regression approaches described here. Diamond [Diamond 88] extends the least squares on fuzzy sets but its least square definition does take into account the qualitative nature of the fuzzy sets. Their definition does not consider the stochastic nature of fuzzy sets and thus we do not consider it a relevant option for regression.

In the Symbolic Data Analysis (SDA) framework [Bock 00, Billard 12], interval-valued data are known as variables with observed values being intervals from the set of real numbers. Thus SDA regression is a type of interval regression with interval-input and interval-output data. Billard and Diday [Billard 00] proposed the first approach to fit a linear regression model to symbolic interval-valued data. Their approach consists of fitting a linear model to the midpoint of the interval valued dataset. The lower and upper bounds of the response variable are predicted by applying this model to the lower and upper bounds of the covariates. They proposed another approach [Billard 02] where two independent linear regression models are built. One model fits with lower side of intervals and the other one fits the upper side. The overall model minimizes $SSE_L + SSE_U$ over the training set where SSE_L and SSE_U denote respectively the sum of squared errors of the lower and the upper model. Lima Neto and De Carvalho [de Lima Neto 08] improved the former approach with another method which is again based on two linear regression models. The first model fits the midpoints of the intervals and the second one is a regression over the ranges, which predicts the dependent variable bounds in a more efficient way. This approach considers the minimization of the sum of the mid-point square error plus the sum of the range square error, and the prediction of the dependent variable is based on the mid-point and range predictions.

Domingues et al. [Domingues 10] have also proposed a robust version of the linear regression method for interval valued-data. Their approach is a center and range approach similar to [Billard 00] and [de Lima Neto 08]. Two innovative features are considered in their work: the predicted values are more robust to the presence of interval valued outliers because they estimate the model parameters by considering heavy-tailed probability error distributions. In their experiments, they use real and simulated symbolic interval datasets to compare their introduced method with Lima Neto and De Carvalho's method [de Lima Neto 08] and they demonstrate that their regression model outperforms the other one.

4.5.2 Methods with a probabilistic point of view

Despite all these contributions to symbolic regression and interval regression models, current approaches view the interval regression problem from an optimization point of view which seeks to minimize a pre-defined criterion and do not consider the probabilistic aspects related to regression models. Therefore, one cannot benefit from statistical inference techniques on

these non-probabilistic models. It makes it impossible to have hypothesis tests or confidence intervals on parameter estimates or prediction. Lima Neto et al. [de Lima Neto 09] were the first to consider a probabilistic view of SDA interval regression. They use bivariate generalized linear models (BGLM) proposed by Iwasaki and Tsubaki [Iwasaki 05] in the context of interval-valued data. Their work includes consideration of some important aspects related to the BGLM and a performance comparison of their approach to that of Billard and Diday [Billard 00] and Lima Neto and De Carvalho's [de Lima Neto 08]. Cattaneo and Wiencierz [Cattaneo 11, Cattaneo 12] introduced Likelihood-based Imprecise Regression (LIR) as a very general theoretical framework for regression analysis with imprecise data. [Cattaneo 12] is a refinement of [Cattaneo 11], where they proposed a robust version of imprecise regression. Their method combines nonparametric likelihood inference with imprecise probability where very weak assumptions are needed and different kinds of uncertainty can be taken into account. The mentioned method is nonparametric, in the sense that no assumption about the error distribution is necessary. They assume that the variables have precise values, but they are imprecisely observed. This imprecision can be observed for the predictors, the response variable or both. Their regression method is linear and based on interval dominance. In another work, Wiencierz and Cattaneo [Wiencierz 13] proposed an algorithm derived from the geometrical properties of LIR results. This algorithm determines the set-valued result of a simple linear regression performed with robust LIR with interval data.
In [Ghasemi Hamed 12c], we extended the Possibilistic KNN regression idea to the probability framework. In this work we propose a KNN interval regression method which finds intervals that for all input instances $x \in \mathcal{X}$ simultaneously contain a β proportion of the response values. We called this problem simultaneous interval regression. This is similar to simultaneous tolerance intervals for regression with a high confidence level $\gamma \approx 1$. We considered the simultaneous interval regression problem for KNN without the homoscedasticity assumption. This work is explained in Section 7.

4.6 Conclusion

In this chapter we have seen the definition of mean regression estimation and quantile regression. We have also seen local linear regression which is a simple effective regression technique. LLR can be used in situations where the function to estimate is complex or where we lack sufficient observations to build a parametric model. Then we have seen quantile regression techniques. These estimators are more robust than least-squares models but as shown by two cross validation examples, they suffer from the crossing quantile effect. Finally, we had a quick look at other regression methods, which give intervals as the output of the response variable. We divided them into two big categories: those which have an optimization view of regression (this is the largest part of the interval regression literature); and those which contain methods which are based on probabilistic assumptions (this part of interval regression literature is based on statistical methods). In my opinion, the latter are more appropriate for the regression problem, which is intrinsically a statistical technique.

Chapter 5

Interval Prediction Methods in Regression

Contents

As pointed out in the previous chapter, regression techniques provide estimates of the conditional mean or quantiles of a real-valued random variable $Y(x)$, being the result of an unknown deterministic function $f(x)$ plus a random noise ε. These models are always built with finite sample size ($n < \inf$), thus the predicted mean or quantile is an estimate of the true unknown conditional mean or quantile of the random variable $Y(x) = f(x) + \varepsilon$. Therefore while dealing with datasets of finite size , we need to make some statistical

inferences. In this chapter, we are interested in finding intervals in regression models which contain a desired proportion of the response variable. The contribution of this chapter is the review and the comparison of different least-squares and quantile regression techniques used to find such intervals. Besides, we take advantage of this chapter to address a misunderstood interval prediction method in the machine learning community. We explain its applications and review its drawbacks.

We choose a fixed regression design where dataset $\mathcal{S} = (x_1, Y(x_1)), \cdots, (x_n, Y(x_n))$ is a random sample. The x_i's are deterministic vectors of observations and $Y(x_i)$ are drawn from the distribution of $Y(x_i)$. These distributions are continuous probability distributions. We always suppose that there is one true mean regression function $f(\cdot)$ with a zero mean error and an unknown variance σ^2. The most practical assumption is the Gaussian homoscedastic error, but it is not mandatory. $\mathcal{S}$ is a finite random sample, so the estimated regression model finds a pair of $(\hat{f}, \hat{\sigma})$; $\hat{f}$ denotes the estimated regression function and $\hat{\sigma}$ is the estimated error standard deviation. This pair is a random vector in the probability space of regression models defined for the underlying regression type (for ex: OLS). Note that in the case of error being not normally distributed, the pair $(\hat{f}, \hat{\sigma})$ does not correctly represent the estimated regression model. Thus we will use the symbol $P_{\mathcal{S}}$ instead of $P_{\hat{f},\hat{\sigma}}$ to refer to a probability distribution where the random vector is the estimated regression model based on the random sample $\mathcal{S}$. We also use the following notation:

- $\mathcal{S} = (x_1, Y(x_1)), \cdots, (x_n, Y(x_n))$: the random sample of regression;
- $f(\cdot)$: the true and unknown regression function;
- $f(x)$: the conditional mean of the response variable for specified combination of the predictors;
- $\hat{f}(\cdot)$: the estimated regression function;
- $\hat{f}(x)$: the estimated regression function at point x;
- ε: the error variable;
- σ^2: the true and unknown variance of the error variable;
- $\hat{\sigma}^2$: the estimated variance of the error variable;
- $\sigma^2_{\hat{f}(x)}$: the variance of $\hat{f}(x)$;
- $\hat{\sigma}^2_{\hat{f}(x)}$: the estimated variance of $\hat{f}(x)$;
- $Y(x)$: the conditional response variable for a given combination of the predictors, $Y(x) = f(x) + \varepsilon$;
- $\chi^2_{p,n}$: the p-quantile of a chi-square distribution with n degrees of freedom;

- Z_p: the p-quantile of a standard normal distribution;
- $t_{p,n}$: the p-quantile of a Student t distribution with n degrees of freedom.

5.1 Conventional techniques

This section describes two conventional interval prediction methods used in least squares models. These methods are asymptotic interval prediction techniques but practitioners tend to use them for a wide class of applications, such as prediction intervals, tolerance intervals and simultaneous tolerance intervals. This is explained further in 5.2. The goal of this section is to see their definitions and discuss their properties.

5.1.1 Conventional Interval prediction

One of the most common interval prediction techniques used in practice is to take $[\hat{f}(x) - Z_{\frac{1-\beta}{2}} SSE^{\frac{1}{2}}, \hat{f}(x) + Z_{1-\frac{1-\beta}{2}} SSE^{\frac{1}{2}}])$ as the interval which contains a β proportion of $Y(x)$'s population,where SSE is the average MSE given by a Leave-One-Out (LOO) or a 10-fold cross validation scheme. One might assume that the intervals expressed below have similar properties to the regression tolerance interval defined in the next section.

$$P\left(Y(x) \in \left[\hat{f}(x) - Z_{\frac{1-\beta}{2}} SSE^{\frac{1}{2}}, \hat{f}(x) + Z_{1-\frac{1-\beta}{2}} SSE^{\frac{1}{2}}\right]\right) = \beta. \tag{5.1}$$

As seen in (4.8), the expected value of SSE is approximately equal to the predictive risk. Thus, based on the bias-variance decomposition:

$$\begin{aligned} E(SSE) &= \text{Average MSE } + \frac{1}{n}\sum_{i=1}^{n} \sigma^2(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{Bias}^2_{\hat{f}(x_i)} + \frac{1}{n}\sum_{i=1}^{n} \sigma^2_{\hat{f}(x_i)} + \frac{1}{n}\sum_{i=1}^{n} \sigma^2(x_i) \end{aligned}$$

and we have:

$$E(SSE) = \text{Average Bias}^2_{\hat{f}(x_i)} + \text{Average } \sigma^2_{\hat{f}(x_i)} + \text{Average } \sigma^2(x_i) \tag{5.2}$$

We assume that:

- the error variance for all x's is constant (homoscedasticity).
- the estimator's variance $\sigma^2_{\hat{f}(x)}$ is constant for all x.
- $\hat{f}(x)$ is an unbiased estimator of $f(x)$.
- the error ε, and $\hat{f}(x)$, are independent and both have normal distributions.

Then, we have:

$$E(SSE) = \sigma^2_{\hat{f}(x)} + \sigma^2$$
$$\hat{f}(x) - \varepsilon \sim \mathcal{N}(f(x), \sigma^2_{\hat{f}(x)} + \sigma^2).$$

Considering the fact that n tends to be large and under the above condition, we can consider SSE as an approximation to the variance of the prediction around any point $SSE \approx \sigma^2_{\hat{f}(x)} + \sigma^2$, which results in (5.3):

$$\frac{\hat{f}(x) - Y(x)}{SSE^{\frac{1}{2}}} \sim \mathcal{N}(0, 1). \tag{5.3}$$

Thus under the above conditions, (5.1) becomes asymptotically valid, but it remains non-applicable for finite sample size datasets. Some practitioners might even think that these intervals contain; a proportion β of the distribution of $Y(x)$ for all values of $x \in \mathcal{X}$. As we will see, this is defined by a simultaneous tolerance interval for regression which is discussed in Section 5.2.4.

5.1.2 Point-wise confidence intervals for the mean function

A mean regression model gives an estimate $\hat{f}(x)$ of the unknown true conditional mean of the response variable $f(x)$. Therefore, one might be interested in obtaining confidence intervals on the true mean function by point-wise confidence intervals.

Definition 16 *Point-wise confidence intervals are intervals that, with a desired level of confidence, are guaranteed to contain the conditional mean regression function $f(x)$.*

$$P_{\hat{f}(x)}(f(x) \in I^{pw}_{1-\alpha}(x)) = 1 - \alpha. \tag{5.4}$$

The estimated function is usually assumed to have an asymptotic normal distribution [Härdle 90, Fan 95]. In case of asymptotic bias, the center of the distribution is shifted and it depends on derivatives of the regression curve (and the distribution of X if we have a random regression model). Given such assumptions, normal point-wise confidence intervals are calculated as follows:

$$I^{pw}_{1-\alpha}(x) = \left(\hat{f}(x) - \widehat{bias}(x)\right) \pm Z_{\frac{\alpha}{2}} \hat{\sigma}_{\hat{f}(x)}, \tag{5.5}$$

where $\widehat{bias}(x)$ is the estimated asymptotic conditional bias. ***The practical assumption is to ignore the bias and assume that the conditional estimated mean function $\hat{f}(x)$ has a constant variance.*** Figure 5.1 shows a simple linear model built on a dataset of 50 observations. The red line represents the true mean function and the blue lines are the different OLS obtained with different random sample generated from the same regression function. In this figure, point-wise 0.95-confidence intervals for the mean function are shown with dashed orange lines.

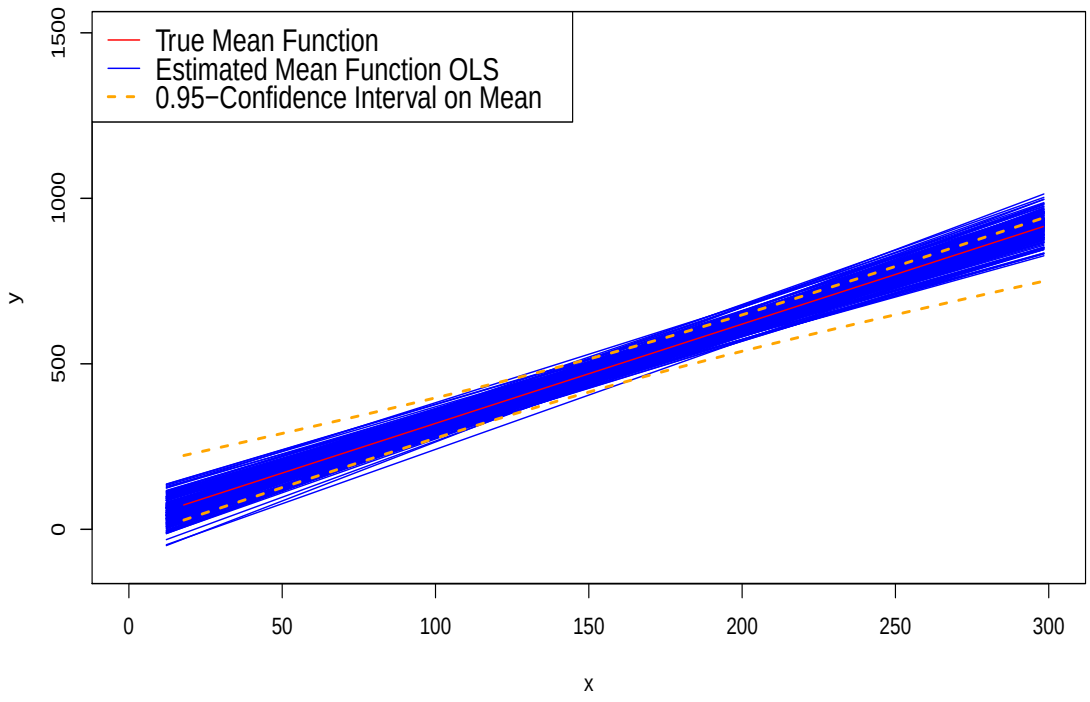

Figure 5.1: Point-wise confidence intervals for the mean function.

Another common method for constructing point-wise confidence intervals uses bootstrapping. This method is applied differently to fixed design and random design. This is because bootstrap the is a re-sampling method that needs bootstrap sample using the same random procedure that has generated the initial dataset. For more details see [Härdle 90] and [Godfrey 09].

5.2 Least-Squares inference techniques

This section describes some well-known statistical inference techniques applied to least-squares regression models. Prediction and tolerance intervals have some equivalents in the quantile regression set-up, but confidence band and simultaneous tolerance intervals seem to be unique to the least-squares world. The goal of this section is to emphasize that least-squares interval prediction methods are not restricted to large sample techniques. There is an extensive literature on this topic. However, there are still some subjects like tolerance intervals and simultaneous intervals that need further study.

5.2.1 Prediction interval for least-squares regression

For a given regression dataset and a query point x, a $1-\alpha$ prediction interval contains, on average, a proportion $(1-\alpha)$ of the distribution of the response variable $Y(x)$. In other words, the expected proportion of the population of $Y(x)$ contained in its $(1-\alpha)$-prediction interval $I_{1-\alpha}^{Pred}(x)$ is $1-\alpha$.

Definition 17 *Let $\mathcal{S} = (x_1, Y(x_1)), \cdots, (x_n, Y(x_n))$ denote a random sample, where the x_is are observations, and $Y(x_i)$ are drawn from the continuous distribution of $Y(x_i)$. A $(1-\alpha)$-prediction regression interval for the point x contains on average a proportion $(1-\alpha)$ of the population of $Y(x)$. The $I_{1-\alpha}^{Pred}(x)$ notation refer to a $(1-\alpha)$-prediction regression interval [Krishnamoorthy 09]. Then, we have:*

$$P_{\mathcal{S},Y(x)}(Y(x) \in I_{1-\alpha}^{Pred}(x)) = 1-\alpha, \text{ where } Y(x) = f(x) + \varepsilon. \tag{5.6}$$

Suppose that we have n independent pairs of random samples as defined below:

$$((\mathcal{S}, \mathcal{T})_1), \cdots, (\mathcal{S}, \mathcal{T})_n)$$

such that:

$$\mathcal{S}_i = (x_1, Y(x_1)), \cdots, (x_k, Y(x_k))_i \text{ and } \mathcal{T}_i = (x_1, Y(x_1)), \cdots, (x_l, Y(x_l))_i,$$

where l and k are arbitrary positive numbers, and we have built a regression model for the first sample of each pair: $\mathcal{S}_i$. If for a given $x = x^*$, one calculates the $(1-\alpha)$-prediction interval of the first sample at point x and then check whether the value(s) of $Y(x^*)$ in the second sample ($\mathcal{T}_i$) are included in the computed $(1-\alpha)$-prediction interval $I_{1-\alpha}^{Pred}(x^*)$, one can observe that, a fraction $1-\alpha$ of $I_{1-\alpha}^{Pred}(x^*)$ will, in the long run, contain the future value(s) of $Y(x^*)$. Note that both the different pairs of samples and the observations within each sample must be independent [Hahn 91].

Regression prediction intervals are also known as $(1-\alpha)$-expectation regression tolerance intervals. An expectation regression tolerance interval is such that its average content is $1-\alpha$. Thus, interval $[L_{1-\alpha}^{Pred}(x), U_{1-\alpha}^{Pred}(x)]$, which is based on a random sample $\mathcal{S}$, is a $(1-\alpha)$-regression prediction interval for observing the next observation of the random variable $Y(x)$ is also a $(1-\alpha)$-expectation tolerance interval.

$$E_{\mathcal{S}}(P(Y(x) \in I_{1-\alpha}^{EXT}(x)|\mathcal{S})) = 1-\alpha, \text{ where } Y(x) = f(x) + \varepsilon. \tag{5.7}$$

For a detailed discussion about the differences between prediction and tolerance intervals, the reader can find more in [Hahn 91, Krishnamoorthy 09, Paulson 43].

Prediction interval in OLS

Suppose that we have an OLS model as explained in 4.2.1, and let x^* be a point in the predictor space which may be previously, observed or not. We know from (4.14) that the estimated vector of parameters $\hat{\beta}$ in OLS has a normal distribution, so the prediction $\hat{y}^* = \hat{f}(x^*)$ is also normally distributed:

$$\hat{y}^* \backsim \mathcal{N}(x^{*T}\beta, \sigma^2 x^{*T}(X^T X)^{-1}x^*), \tag{5.8}$$

and we can conclude that:

$$\frac{(y - \hat{y}^*)}{\sqrt{\frac{n\hat{\sigma}^2}{n-p}(1 + x^{*T}(X^T X)^{-1}x^*)}} \backsim t_{n-p}. \tag{5.9}$$

So we have:

$$\hat{y}^* - \mathbf{c}t_{(1-\frac{\alpha}{2},n-p)} \leq y \leq \hat{y}^* + \mathbf{c}t_{(1-\frac{\alpha}{2},n-p)}, \tag{5.10}$$

$$\mathbf{c} = \sqrt{\frac{n\hat{\sigma}^2}{n-p}(1 + x^{*T}(X^T X)^{-1}x^*)}$$

and (5.10) gives a two tailed $1-\alpha$ prediction interval for the OLS [Mendenhall 06]. These intervals are illustrated in Figure 5.2.

Bonferroni Prediction intervals

Prediction intervals are confidence intervals for the response variable $Y(x) = f(x) + \varepsilon$. They can be also constructed by using the Bonferroni inequality to construct simultaneous confidence statements on both the mean regression function and the error variable. Let the intervals $I^{pw}_{1-\frac{\alpha}{2}}(f(x)) = [L^{pw}_{1-\frac{\alpha}{2}}(f(x)), U^{pw}_{1-\frac{\alpha}{2}}(f(x))]$ and $I^{C}_{1-\frac{\alpha}{2}}(\varepsilon) = [L^{C}_{1-\frac{\alpha}{2}}(\varepsilon), U^{C}_{1-\frac{\alpha}{2}}(\varepsilon)]$ be respectively a $(1-\frac{\alpha}{2})$ point-wise confidence interval for the conditional mean at point x and the confidence interval for error variable such that:

$$P_{\hat{f}}\left(f(x) \in I^{pw}_{1-\frac{\alpha}{2}}(f(x))\right) = 1 - \frac{\alpha}{2}, \tag{5.11}$$

and

$$P_{\hat{\sigma},\varepsilon}\left(\varepsilon \in I^{C}_{1-\frac{\alpha}{2}}(\varepsilon)\right) = 1 - \frac{\alpha}{2}.$$

Then based on the Bonferroni inequality we have:

$$P_{\hat{f},\hat{\sigma},\varepsilon}\left(\left[f(x) \in I^{C}_{1-\frac{\alpha}{2}}(f(x))\right] \text{ and } \left[\varepsilon \in I^{C}_{1-\frac{\alpha}{2}}(\varepsilon)\right]\right) = 1 - \alpha.$$

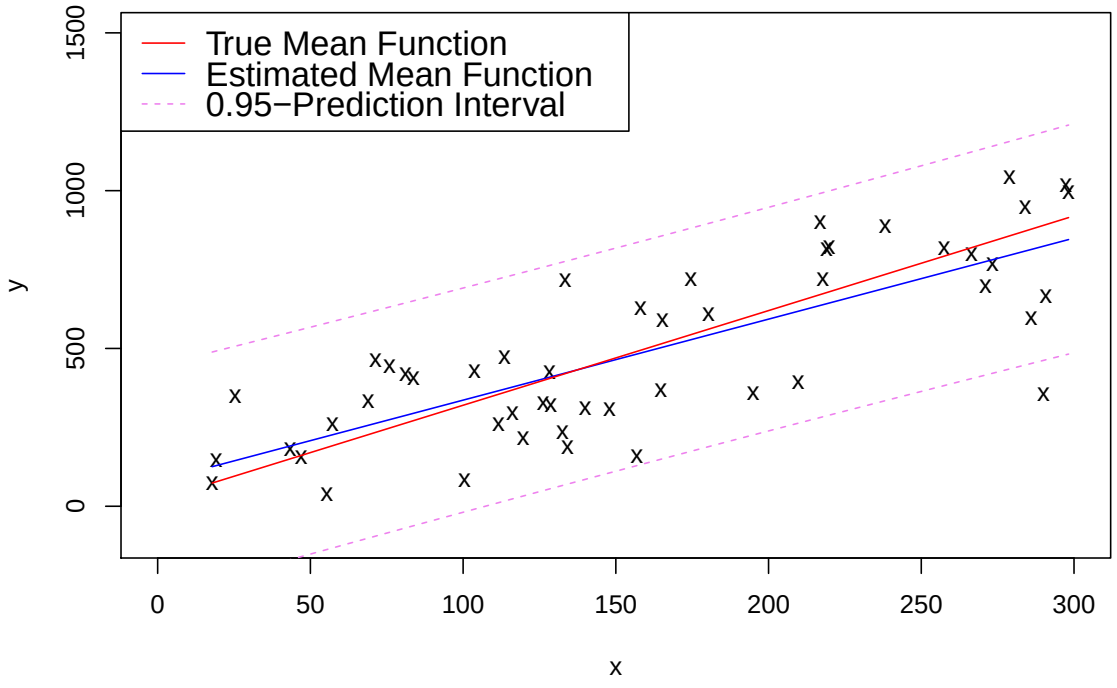

Figure 5.2: Predition intervals for OLS.

Therefore we have (5.12), which describes a prediction interval for the regression function.

$$P_{\hat{f},\hat{\sigma},\varepsilon}\Big(Y(x) \in I^{Pred}_{1-\alpha}(x)\Big) = 1-\alpha, \text{ where } I^{Pred}_{1-\alpha}(x) = [L^{pw}_{1-\frac{\alpha}{2}}(f(x)) + L^{C}_{1-\frac{\alpha}{2}}(\varepsilon), U^{pw}_{1-\frac{\alpha}{2}}(f(z)) + U^{C}_{1-\frac{\alpha}{2}}(\varepsilon)]. \tag{5.12}$$

5.2.2 Confidence bands for least-squares regression

Confidence bands are simultaneous point-wise confidence intervals. The idea is to have intervals that, with confidence level $1-\alpha$, contain the entire true mean function.

Definition 18 *Confidence bands are random intervals $[U_{1-\alpha}(x), L_{1-\alpha}(x)]$ that include, with probability $1-\alpha$, the entire true mean regression function $f(\cdot)$.*

$$P_{\hat{f}}\Big(f(x) \in I^{cb}_{1-\alpha}(x) \text{ for all } x \in \mathcal{X}\Big) = 1-\alpha, \text{ where } I^{cb}_{1-\alpha}(x) = [U_{1-\alpha}(x), L_{1-\alpha}(x)], \tag{5.13}$$

where $\mathcal{X}$ is the domain of x.

Confidence bands are usually hard to compute, even for parametric models. Working and Hotelling [Working 29] were the first to propose the confidence band for the simple linear regression and Scheffe [Scheffé 59] generalized it to the linear regression with multiple predictors. Sun and Loader [Sun 94] generalized the linear case to the non-linear regression with linear estimates. They provided an approximation to the tube formula which can be used for multidimensional predictors and a wide class of linear estimates. There are three general ways to compute non-linear confidence bands [Härdle 90]:

- Bonferroni approach: confidence bands are simultaneous confidence intervals around the mean function. Therefore one common way to obtain uniform confidence bands is to use point-wise confidence intervals with a confidence level adjusted by the Bonferroni inequality. Some authors have already studied this approach [Eubank 93, De Brabanter 11]. It is well suited if the band is required for a small number of points otherwise it provides wide intervals.

- Gaussian process approximation: this approach consists of considering $\hat{f}(x) - f(x)$ as a Gaussian process and deriving its asymptotic Gaussian process approximation [Sun 94]. For this purpose we need the distribution of the maximum of a Gaussian process and fortunately this is a well studied problem.

- Bootstrap method: this technique uses re-sampling in order to approximate the distribution of the maximum deviation from the mean

$$W_n = \sup_{x \in \mathbf{X}} |\hat{f}(x) - f(x)|.$$

  The lower and upper bands will be the $\frac{\alpha}{2}$-quantile and $(1-\frac{\alpha}{2})$-quantile of W_n. Another approach consists of approximating the distribution of $\hat{f}(x) - f(x)$ for each x and then correcting the confidence level of all points in order to have simultaneous coverage of $1-\alpha$ [Hardle 91, Härdle 90].

Figure 5.3 use orange solid lines to represent the Working and Hotelling confidence band. We can see that they are larger than point-wise confidence intervals for the mean function.

5.2.3 Tolerance intervals for least-squares regression

In the case of regression with constant error variance and normal distribution of errors, usually inter-quantiles of a normal distribution with mean zero and variance $\hat{\sigma}^2$, (being the error variance estimator) are used as an approximate solution to find intervals that contain a desired proportion of the distribution of the response variable for a given value of dependent variables. For instance, the 0.95 inter-quantile $[\hat{f}(x) - 1.96\hat{\sigma}, \hat{f}(x) + 1.96\hat{\sigma}]$ is often used as the interval containing 95% of the distribution of $Y(x)$ (i.e., as a regression tolerance interval). As shown by Wallis [Wallis 51], this statement is not true since $\hat{\sigma}^2$ and $\hat{f}(x)$ are only estimations of the true error variance σ^2 and the true mean function at point x, $f(x)$. These estimations are always made on a finite sample and are then pervaded with

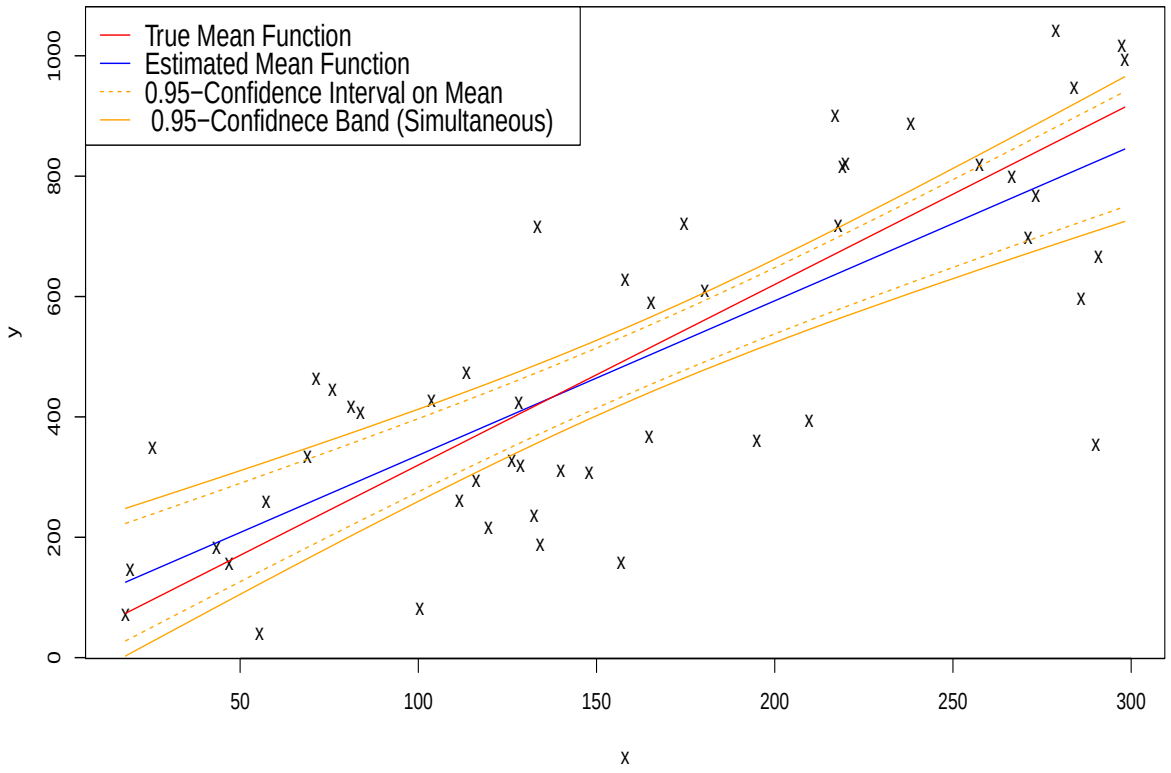

Figure 5.3: Working and Hotelling confidence band in OLS for a random sample with $n = 50$.

uncertainty. Tolerance intervals for least squares regression have been introduced in order to take into account this uncertainty. These intervals are described formally by (5.14). We will refer to such intervals, β-content γ-coverage regression tolerance intervals and they are denoted by $I^T_{\gamma,\beta}(x)$.

$$P\left(\int_{U^T_{\beta,\gamma}(x)}^{U^T_{\beta,\gamma}(x)} p_x(t)dt \geq \beta\right) = \gamma \text{ where } Y(x) = f(x) + \varepsilon, \tag{5.14}$$

where $p_x(t)$ denotes the probability density function of $Y(x)$ for a specified value of the predictor variable x. A two-sided tolerance interval $I^T_{\gamma,\beta}(x)$ for $Y(x)$ is taken, of the form $\hat{f}(x) \pm \rho(x)\hat{\sigma}$, where $\rho(x)$ is the tolerance factor to be determined subject to the content β and the desired confidence level γ. Let $C(x; \hat{f}, \hat{\sigma})$ represent the content of this tolerance

interval,

$$C(x;\hat{f},\hat{\sigma}) = P_{Y(x)}\Big(\hat{f}(x) - \rho(x)\hat{\sigma} \leq Y(x) \leq \hat{f}(x) + \rho(x)\hat{\sigma}\Big). \tag{5.15}$$

The tolerance factor $\rho(x)$ satisfies the following condition:

$$P_{\hat{f},\hat{\sigma}}\Big(C(x;\hat{f},\hat{\sigma}) \geq \beta\Big) = \gamma. \tag{5.16}$$

Equations (5.14) and (5.16) could also be expressed as follows:

$$P_{\hat{f},\hat{\sigma}}\Big(P_{Y(x)}\Big(Y(x) \in I^T_{\gamma,\beta}(x)\Big) \geq \beta\Big) = \gamma, \tag{5.17}$$

$$I^T_{\gamma,\beta}(x) = [L^T_{\beta,\gamma}(x), U^T_{\beta,\gamma}(x)] = [\hat{f}(x) - \rho(x)\hat{\sigma}, \hat{f}(x) + \rho(x)\hat{\sigma}].$$

It is important to observe that tolerance intervals in regression are defined separately for each input vector. Therefore, for two different input vectors $x = x_1$ and $x = x_2$, $I^T_{\gamma,\beta}(x_1)$ and $I^T_{\gamma,\beta}(x_2)$ are different and the event $Y(x_1) \in I^T_{\gamma,\beta}(x_1)$ is independent of $Y(x_2) \in I^T_{\gamma,\beta}(x_2)$. For more details see [Hahn 91] and [Krishnamoorthy 09].

Bonferroni regression tolerance intervals

Tolerance intervals for regression can be also constructed by using the Bonferroni inequality. In this part, we demonstrate how to find Bonferroni regression tolerance intervals for the OLS. We have chosen the OLS model because we have already described all the formulae, and it is the simplest case but this technique can be applied to any regression model. For this purpose one must use the Bonferroni inequality in order to combine the confidence bands on the regression's mean and the confidence interval on the error's standard deviation.

The first part is exactly the same as the first step in Bonferroni Prediction intervals described in 5.2.1. It consists of finding a point-wise confidence interval for the conditional mean at point x described by (5.11). In the next step, an upper bound on the error's standard deviation must be obtained. In the case of OLS, we use (4.16) which results in (5.18) where N denotes the number of observations, and k the number of predictors in the OLS model. In non-linear models, the error's standard deviation, are estimated by other methods and more details can be found in [Gasser 86, Yu 04].

$$P_{\hat{\sigma}}(\sigma \leq \mathbf{c}\hat{\sigma}) = 1 - \frac{\alpha}{2}, \text{ where } \mathbf{c} = \left(\frac{N-k-1}{\chi^2_{\frac{\alpha}{2},N-k-1}}\right)^{\frac{1}{2}}. \tag{5.18}$$

Now by applying the Bonferroni inequality, one can use the confidence statements (5.18) and (5.11) to obtain a joint confidence statement with probability greater or equal than

$1-\alpha$. Equation (5.19) describes this combination.

$$P_{\hat{f},\hat{\sigma}}\left(\left[P_{\varepsilon}\left(Z_{\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma} \leq \varepsilon \leq Z_{1-\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma}\right) = \beta\right] \text{ and } \left[f(x) \in I^{pw}_{1-\frac{\alpha}{2}}(f(x))\right]\right) \geq 1-\alpha, \quad (5.19)$$

$$\text{where } \mathbf{c} = \left(\frac{N-k-1}{\chi^2_{\frac{\alpha}{2},N-k-1}}\right)^{\frac{1}{2}}, I^{pw}_{1-\frac{\alpha}{2}}(f(x)) = [L^{pw}_{1-\frac{\alpha}{2}}(f(x)), U^{pw}_{1-\frac{\alpha}{2}}(f(x))].$$

By rewriting the statement above, we find tolerance intervals for a regression function.

$$P_{\hat{f},\hat{\sigma}}\left(P_{\varepsilon}\left(L^{pw}_{1-\frac{\alpha}{2}}(f(x)) + Z_{\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma} \leq f(x)+\varepsilon \leq U^{pw}_{1-\frac{\alpha}{2}}(f(x)) + Z_{1-\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma}\right) = \beta\right) \geq \gamma, \quad (5.20)$$

$$\text{where } \mathbf{c} = \left(\frac{N-k-1}{\chi^2_{\frac{\alpha}{2},N-k-1}}\right)^{\frac{1}{2}}, \gamma = 1-\alpha.$$

Figure 5.4 represent the Bonferroni regression tolerance intervals in OLS. We can see that they are larger than prediction intervals.

5.2.4 Simultaneous tolerance intervals for least-squares regression

As seen above, tolerance intervals for least squares regression are point-wise intervals which are obtained separately for each vector of x. Lieberman and Miller [Lieberman 63] extended the Wallis [Wallis 51] idea to the simultaneous case. Simultaneous tolerance intervals are constructed in such a way, that with confidence level γ, simultaneously for all possible values of input vector x, at least a proportion β of the whole population of the response variable Y is contained in the obtained intervals. Simultaneous tolerance intervals for least squares regression $[L^{TS}_{\beta,\gamma}(x), U^{TS}_{\beta,\gamma}(x)]$ create an envelope around the entire mean regression function $f(\cdot)$ such that, for all $x \in \mathcal{X}$, the probability that $Y(x)$ is contained in $[L^{TS}_{\beta,\gamma}(x), U^{TS}_{\beta,\gamma}(x)]$ is simultaneously β, and this coverage is guaranteed with a confidence level γ. We name such intervals β-content γ-coverage simultaneous regression tolerance intervals. We represent them by $I^{TS}_{\gamma,\beta}(x)$ and they are described by (5.21), where $p_x(t)$ represents the probability density function of $Y(x)$ for a specified value of the predictor variable x.

$$P\left(\min_{x\in\mathcal{X}}\left(\int_{L^{TS}_{\beta,\gamma}(x)}^{U^{TS}_{\beta,\gamma}(x)} p_x(t)dt\right) \geq \beta\right) = \gamma, \text{ where } Y(x) = f(x)+\varepsilon. \quad (5.21)$$

If $\rho(x)$ in (5.15) is a simultaneous tolerance factor, then it must satisfy the following condition:

$$P_{\hat{f},\hat{\sigma}}\left(\min_{x\in\mathcal{X}} C(x;\hat{f},\hat{\sigma}) \geq \beta\right) = \gamma. \quad (5.22)$$

These intervals have been studied for the linear regression by several authors [Lieberman 63, Wilson 67, Mee 91]. For an introduction to the subject, the reader can see Lieberman and Miller [Lieberman 63]. They explained the problem in detail and presented four different methods for construction of such intervals for linear regression. For more information about simultaneous inference, see [Krishnamoorthy 09, Miller 91].

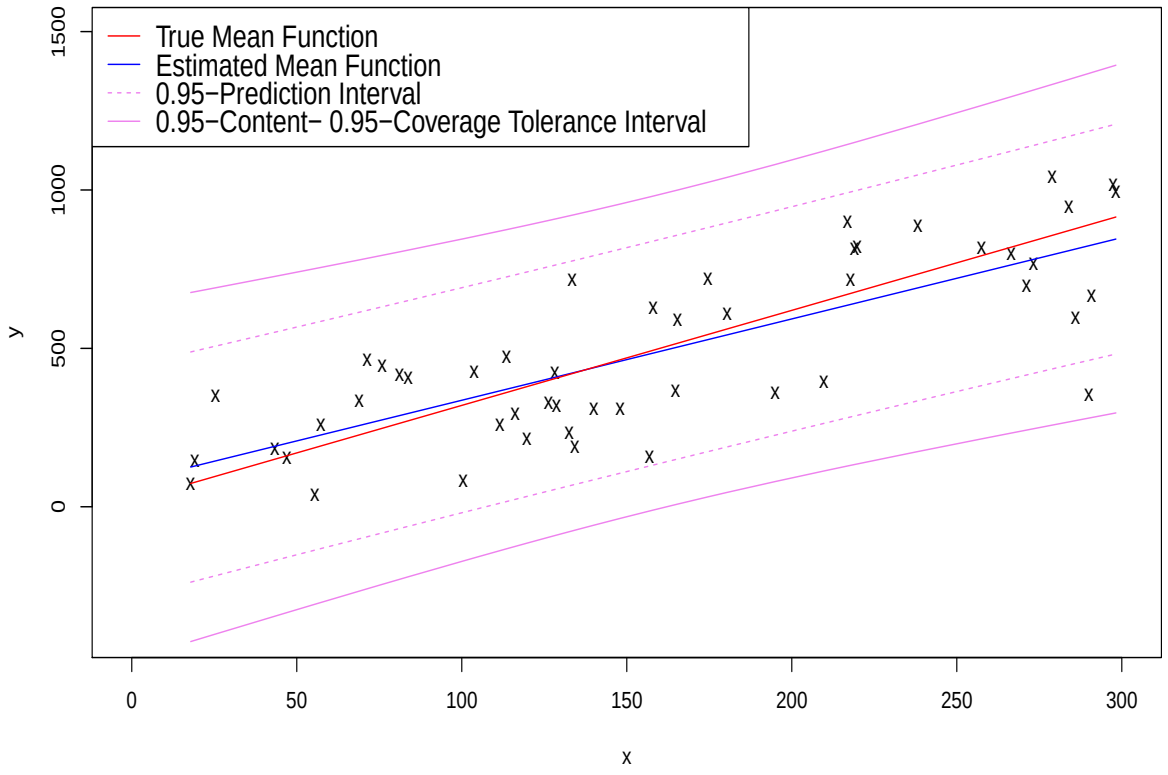

Figure 5.4: Bonferroni regression tolerance intervals in OLS for a random sample with $n = 50$.

Bonferroni Simultaneous tolerance intervals

Lieberman and Miller [Lieberman 63] used the Bonferroni inequality to construct simultaneous tolerance intervals for linear regression. However, this approach can be applied in other models with a constant and normal error variance. For this purpose one must use the Bonferroni inequality in order to combine the confidence bands on the regression mean and the confidence interval on the error standard deviation.

As explained in Section (5.2.2), a $(1 - \frac{\alpha}{2})$-confidence band for the mean regression function is a tube that, with confidence level $(1 - \frac{\alpha}{2})$, contains the entire mean function $f(\cdot)$ and (5.23) describes such bands.

$$P_{\hat{f}}\left(f(x) \in I^{cb}_{1-\frac{\alpha}{2}}(x) \text{ for all } x \in \mathcal{X}\right) = 1 - \frac{\alpha}{2}, \text{ where } I^{cb}_{1-\frac{\alpha}{2}}(x) = [L^{cb}_{1-\frac{\alpha}{2}}(x), U^{cb}_{1-\frac{\alpha}{2}}(x)]. \quad (5.23)$$

Confidence bands for the linear regression can easily be obtained with Scheffe's technique

[Scheffé 59], but it becomes harder for non-linear models. Then we must obtain an upper bound on the error's standard deviation. In the case of OLS, we can use (5.24), where N denotes the number of observations, k the number of predictors in the OLS model and it is a direct result of (4.16). In non-linear models the error's standard deviation, are estimated by other methods and the interested reader can find details in [Gasser 86, Yu 04].

$$P_{\hat{\sigma}}(\sigma \leq \mathbf{c}\hat{\sigma}) = 1 - \frac{\alpha}{2}, \text{ where } \mathbf{c} = \left(\frac{N-k-1}{\chi^2_{\frac{\alpha}{2},N-k-1}}\right)^{\frac{1}{2}}. \tag{5.24}$$

Now by applying the Bonferroni inequality, one can use the confidence statements (5.24) and (5.23) to obtain a joint confidence statement with probability greater or equal to $1-\alpha$. Equation (5.25) describes this combination.

$$P_{\hat{f},\hat{\sigma}}\left(\left[P_{\varepsilon}\left(Z_{\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma} \leq \varepsilon \leq Z_{1-\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma}\right) = \beta\right] \text{ and } \left[f(x) \in I^{cb}_{1-\alpha}(x) \text{ for all } x \in \mathcal{X}\right]\right) \geq 1-\alpha, \tag{5.25}$$

$$\text{where } \mathbf{c} = \left(\frac{N-k-1}{\chi^2_{\frac{\alpha}{2},N-k-1}}\right)^{\frac{1}{2}}, I^{cb}_{1-\frac{\alpha}{2}}(x) = [L^{cb}_{1-\frac{\alpha}{2}}(x), U^{cb}_{1-\frac{\alpha}{2}}](x).$$

By rewriting the above statement, we find simultaneous tolerance intervals for a regression function:

$$P_{\hat{f},\hat{\sigma}}\left(P_{\varepsilon}\left(L^{cb}_{1-\frac{\alpha}{2}}(x) + Z_{\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma} \leq f(x) + \varepsilon \leq U^{cb}_{1-\frac{\alpha}{2}}(x) + Z_{1-\frac{1-\beta}{2}}\mathbf{c}\hat{\sigma}\right) = \beta, \text{ for all } x \in \mathcal{X}\right) \geq \gamma, \tag{5.26}$$

$$\text{where } \mathbf{c} = \left(\frac{N-k-1}{\chi^2_{\frac{\alpha}{2},N-k-1}}\right)^{\frac{1}{2}}, \gamma = 1-\alpha.$$

Lieberman and Miller [Lieberman 63] compared this approach to other simultaneous tolerance intervals for simple linear regression. These resulting intervals have the nominal β-content but they tend to be too wide. Figure 5.5 represent the these intervals in OLS.

5.3 Interval prediction with Quantile Regression Models

While the least-squares models just estimates the conditional mean function $f(x)$, the quantile regression obtains estimates of conditional quantiles which, on average, estimates the true quantile function. A quantile regression model can estimate one conditional quantile, so with quantile regression models, we can find one-sided intervals:

$$IQ_{1-\alpha}(x) = (-\infty, Q_{1-\alpha}(x)], \tag{5.27}$$

where $Q_{1-\alpha}(x)$ estimates the true $(1-\alpha)$-quantile of the conditional distribution of the response value $Y(x)$, given a particular combination of predictors. Note that as seen in

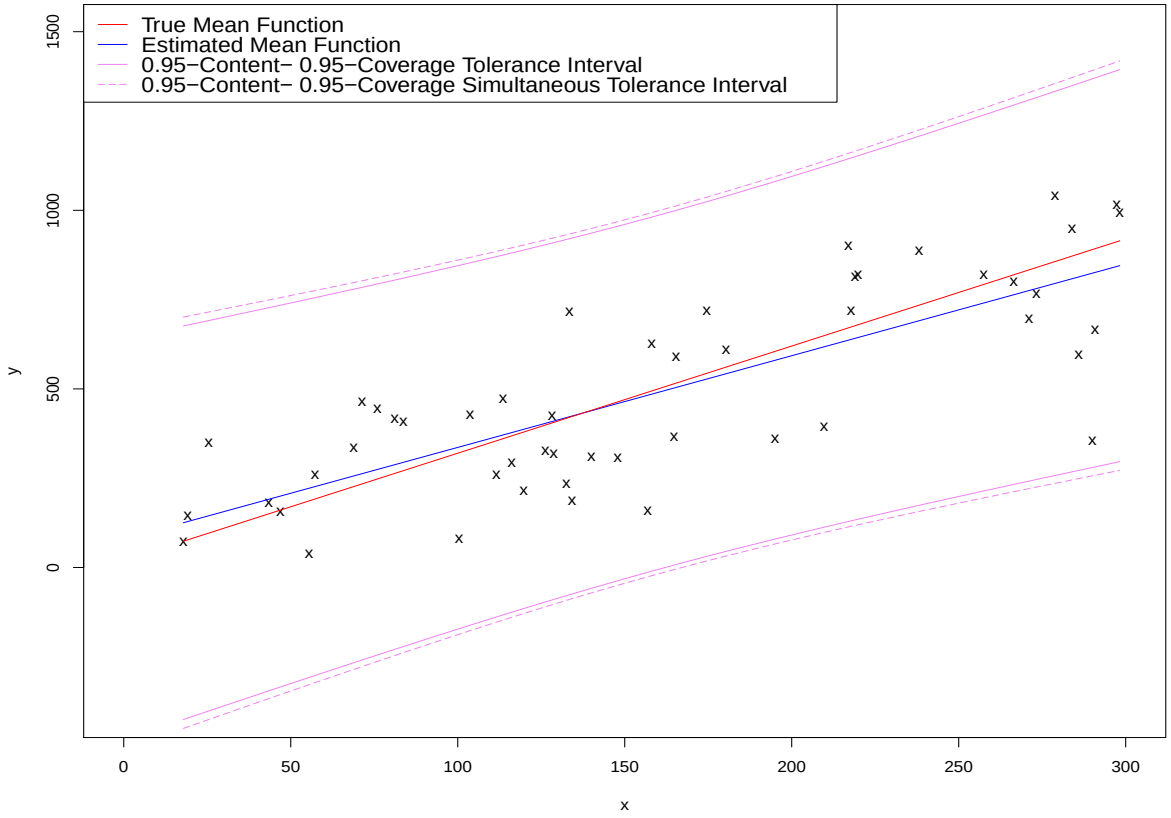

Figure 5.5: Bonferroni tolerance intervals and Bonferroni simultaneous tolerance intervals in OLS for a random sample with $n = 50$.

Section 4.4, $Q_{1-\alpha}(x)$ is just an estimation of the unknown conditional quantile function. This means that interval $IQ_{1-\alpha}(x)$ will, on average, contain a proportion $1 - \alpha$ of $Y(x)$. In the following section, we will have a quick look at different ways of obtaining confidence intervals on regression quantiles and then we will see how quantile regression can be used to predict one-sided and two-sided conditional intervals.

5.3.1 Confidence interval on regression quantiles

Once we have estimated our conditional quantile with a quantile regression model, we need a method of statistical inference to obtain confidence intervals on the conditional quantile. Equations (4.36) and (4.37) give respectively the covariance matrix and the variance of $\hat{\beta}_\tau$. However, evaluating these quantities require the value of the error's density function at the origin ($\frac{1}{f(0)}$), called the sparsity, which is itself unknown. Koenker [Koenker 94a] studied some methods for estimating $\frac{1}{f(0)}$. In the same study, he used a Monte-Carlo experiment to

evaluate the discussed methods, and this study showed that confidence intervals based on (4.37) are not robust for heteroscedastic error models.

There are already three classes of methods to find these confidence intervals [Kocherginsky 05].

- Direct estimation: this method uses kernel-smoothing to obtain direct estimations of the sparsity for the iid and non-iid error model. This method holds generally when the sample size increases and the bandwidth of the smoother tends to zero, but for finite sample size its performance depends strongly on the choice of the smoother's bandwidth.

- Rank-Score Method: Koenker [Koenker 94a] proposed another approach for finding confidence intervals on linear regression quantiles. This method is based on inverting a regression rank score test. This method has reliable results and is robust to deviations from the model, but it is practically feasible only for one-dimensional parameter and becomes computationally prohibitive for large datasets and does not provide any estimate of the covariance matrix [Kocherginsky 05].

- Re-sampling Method: the two common methods are bootstrapping residuals (holds only for iid models) and bootstrapping pairs. Pairwise bootstrap is a rather effective method to estimate the confidence intervals. These methods require repeated calculation of regression quantiles, which can become very time consuming when n and p increase. He and Hu. [He 02] proposed a new re-sampling method based on the Markov Chain Marginal Bootstrap (MCMB). Instead of solving a p-dimensional system for each bootstrap replicate, as is usually done in the bootstrap, it completes each bootstrap replication by solving p one-dimensional equations. This method decreases the computational complexity associated with bootstrap application in high-dimensional spaces. *Note that this method is only proposed for the linear models in the class of quantile regression estimators.*

Kocherginsky et al. [Kocherginsky 05] compared several methods of the three classes listed above and they proposed different approaches depending on the size of the dataset and the number of variables. Kocherginsky et al.[Kocherginsky 05] have also proposed a simple modification to the He and Hu [He 02] MCMB method which yields a more time-saving method for obtaining quantile regression confidence intervals. ***Note that all the experiences and results on confidence intervals for regression quantiles have so far only been done for linear models.***

5.3.2 One-sided interval prediction

In quantile regression models, if we want one-sided intervals that contain a desired proportion of the conditional distribution of the response variable, we have two choices:

- **Estimates of point-wise interval:** This is the definition of quantile regression explained in 4.4.1. The estimated intervals are defined by (5.27). These intervals

contain, on average, a proportion $1-\alpha$ of the conditional distribution of the response variable $Y(x)$. These interval are similar to one-sided prediction intervals in least-squares models defined in 5.2.1.

- **Confidence based point-wise inference:** This is the definition of the confidence interval on regression quantiles explained in 5.3.1. This one-sided upper γ confidence interval on $(1-\alpha)$-regression quantile will be of the form:

$$IQ^{\gamma}_{1-\alpha}(x) = (-\infty, Q^{\gamma}_{1-\alpha}(x)], \tag{5.28}$$

where $Q^{\gamma}_{1-\alpha}(x)$ is an upper or lower γ-confidence bound on the conditional quantile of $Y(x)$.

  - An upper γ-confidence interval on the $1-\alpha$ quantile of $Y(x)$ will cover, the $1-\alpha$ quantile of $Y(x)$ *with at least a proportion* γ. This is described by $IU^{\gamma}_{1-\alpha}(x)$:

$$IU^{\gamma}_{1-\alpha}(x) = (-\infty, U^{\gamma}_{1-\alpha}(x)],$$

    where $U^{\gamma}_{1-\alpha}(x)$ is the upper confidence bound.
  - In contrast, a lower γ-confidence interval on the $1-\alpha$ quantile of $Y(x)$ will cover, the $1-\alpha$ quantile of $Y(x)$ *with at most a proportion* γ. This is described by $IL^{\gamma}_{1-\alpha}(x)$:

$$IL^{\gamma}_{1-\alpha}(x) = (-\infty, L^{\gamma}_{1-\alpha}(x)], \tag{5.29}$$

    where $L^{\gamma}_{1-\alpha}(x)$ is the upper confidence bound.

Note that by using lower confidence bounds, we can obtain intervals $I^{\gamma}_{\alpha}(x)$ such that:

$$I^{\gamma}_{\alpha}(x) = [L^{\gamma}_{1-\alpha}(x), \infty).$$

$I^{\gamma}_{\alpha}(x)$ is an upper γ-confidence interval that covers, *at least* γ *of the time*, an interval containing the $1-\alpha$ proportion of $Y(x)$. *As stated in 5.3.1 all the experiments and results on confidence interval for regression quantiles have so far only been performed for linear models.* These intervals are similar to one-sided tolerance intervals for regression in least-squares models explained in 5.2.3.

5.3.3 Two-sided interval prediction

In order to obtain two-sided $(1-\alpha)$-content conditional intervals, one must build two distinct quantile regression models: a lower $\frac{\alpha}{2}$-quantile regression model and an upper $(1-\frac{\alpha}{2})$-quantile regression model. As in one-sided conditional intervals, we have two choices:

- **Estimates of point-wise interval:** One must build two distinct quantile regression models to obtain two-sided intervals. Thus we build a pair of models which consist of a lower $(\frac{\alpha}{2})$-quantile regression model and an upper $(1-\frac{\alpha}{2})$-quantile regression model. For example, in order to obtain 90-predictive intervals with Linear Quantile Regression (LQR), we construct a lower 0.05-LQR model and an upper 0.95-LQR model. Such pairs of models can provide two-sided intervals which contain, on average, a desired proportion $1-\alpha$ of the distribution of $Y(x)$. These intervals are similar to two-sided prediction intervals in least-squares models defined in 5.2.1.

- **Confidence based point-wise inference:** These two-sided intervals contain with a γ confidence level, a proportion $1-\alpha$ of $Y(x)$. As described above, we need a pair of $(\frac{\alpha}{2}, 1-\frac{\alpha}{2})$ quantile regression models but each model now needs itself a confidence interval as explained in 4.4.1. Suppose that we have built the upper and lower quantile regression models and let $\gamma = 1-\tau$, now we must obtain a lower (one-sided) $(1-\frac{\tau}{2})$ confidence interval on the lower $\frac{\alpha}{2}$-quantile regression model and an upper (one-sided) $(1-\frac{\tau}{2})$ confidence interval on the upper $(1-\frac{\alpha}{2})$-quantile regression model. The lower and upper $(1-\frac{\tau}{2})$-confidence intervals are respectively denoted $IL_{\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)$ and $IU_{1-\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)$ in Equations (5.30) and (5.31).

$$P_{\mathcal{S}}\left(P_{Y(x)}(Y(x) \in IL_{\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)|\mathcal{S}) \leq \frac{\alpha}{2}\right) = 1-\frac{\tau}{2}, \text{ where } IL_{\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x) =]-\infty, L_{\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)], \tag{5.30}$$

$$P_{\mathcal{S}}\left(P_{Y(x)}(Y(x) \in IU_{1-\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)|\mathcal{S}) \geq 1-\frac{\alpha}{2}\right) = 1-\frac{\tau}{2}, \text{ where } IU_{1-\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x) =]-\infty, U_{1-\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)]. \tag{5.31}$$

In Equations (5.30) and (5.31), $L_{\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)$ denotes an lower confidence bound on the $\frac{\alpha}{2}$-regression quantile at point x. This confidence bound must, a proportion $1-\frac{\tau}{2}$ of the time, cover the $\frac{\alpha}{2}$ quantile of $Y(x)$. $L_{1-\frac{\alpha}{2}}^{1-\frac{\tau}{2}}(x)$ denotes an upper confidence bound on the regression quantile at point x and it must, a proportion $1-\frac{\tau}{2}$ of the time, cover the $1-\frac{\alpha}{2}$ quantile of $Y(x)$.

Note that these confidence statements are made on two different models, and so we cannot use them directly to construct two-sided confidence intervals. However, by applying the Bonferroni inequality, one can merge the pair of $(1-\frac{\tau}{2})$ confidence intervals to obtain a joint confidence statement with a probability greater than or equal to $\gamma = 1-\tau$. Equation (5.32) describes this combination.

$$P_{\mathcal{S}}\left(P_{Y(x)}\left(Y(x) \in IQ_{1-\alpha}^{1-\tau}(x)\right) \geq 1-\alpha\right) \geq 1-\tau, \tag{5.32}$$

where $IQ^{1-\tau}_{1-\alpha}(x) = [L^{1-\frac{\tau}{2}}_{\frac{\alpha}{2}}(x), U^{1-\frac{\tau}{2}}_{1-\frac{\alpha}{2}}(x)]$.

It is important to emphasize that although these intervals are theoretically feasible, there has not been any work, until now, which treats the problem of two-sided interval prediction with one-sided confidence intervals on regression quantiles. Such intervals are similar to two-sided γ-coverage $1-\alpha$-content least-squares tolerance intervals and they are explained in 5.2.3.

Figure 5.6 shows these intervals for linear model built on a dataset of 50 observations. The red line represents the true mean function and the green lines are the different linear estimate of quantile function obtained with different samples of the same function. We used $0.025-LQR$, $0.975-LQR$, and 0.95-coverage 0.95-content two-sided Bonferroni LQRC to denote respectively linear estimation of 0.025-quantiles, linear estimation of 0.975 quantiles and confidence intervals on regression quantiles combined with the Bonferroni method as explained in Equations (5.30) and (5.31). These lines are shown by orange color and they are obtained for one pair of green lines. They are obtained with the re-sampling method as explained in [Kocherginsky 05]. quantile function. obtained with different random sample generated from the same regression function. In this figure, point-wise 0.95-confidence intervals for the mean function are shown with dashed orange lines.

Note as discussed in 4.4, two different quantile regression models may cross or overlap each other, which is called as *quantile crossing*. Thus two-sided interval prediction is more meaningful by enforcing the *non-crossing* constraint. *However after enforcing this constraint the conditional quantile estimator may not converge to the true conditional quantile. Thus we have to choose between a "non-correct" or non-convergent estimator.*

5.4 Discussion

In the previous chapter we saw different regression techniques. Apart from interval regression methods, the presented techniques are employed to estimate a point of the conditional distribution distribution of the response variable. This point can be a conditional quantile or the conditional mean. This chapter discussed several methods of finding intervals in regression models which contain a desired proportion of the response variable. One common category contains methods that consist of building a least squares or mean estimating regression model and then employing some kind of statistical inference techniques to predict such intervals. Another classical method uses quantile regression models [Koenker 05]. In this section we will give a brief survey of the presented interval prediction methods. This classification is summarized in Figure 5.4. First, we begin by least-squares methods and then we will survey quantile regression interval prediction techniques.

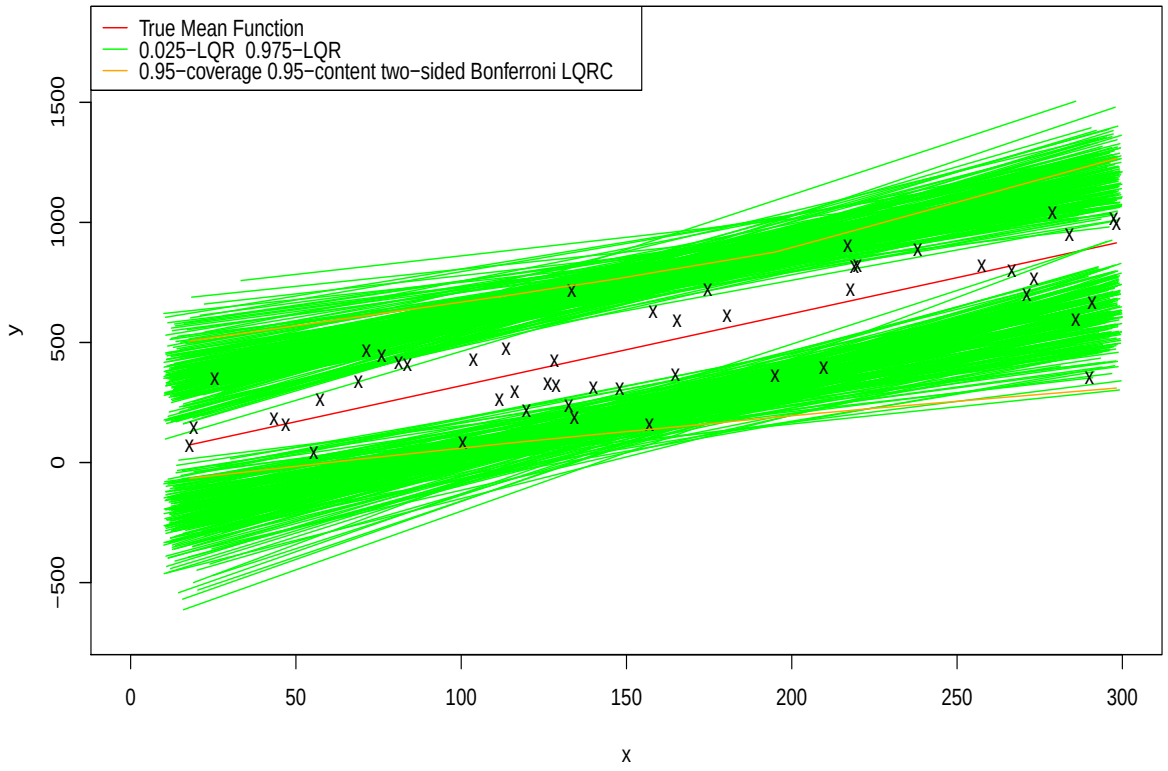

Figure 5.6: Two-sided Bonferroni method for confidence intervals on regerssion quantiles.

5.4.1 Least-Squares models

- **Conventional method**: Some practitioners of the Machine Learning community tend to use a common method to estimate different types on intervals explained below:

  1. The error variance ($\hat{\sigma}^2_{error}$) is just estimated by the training set's average mean squared error usually obtained with a cross-validation schema. This method is explained in Section 5.1.1.
  2. The desired β-inter-quantile of the error's distribution is estimated.
  3. Finally, the estimated inter-quantile is added to $\hat{f}(x)$ (the estimated mean function) and is used to find the inter-quantile of the conditional distribution of response variable $Y(x)$.

  Unfortunately, most practitioners of the Machine Learning community usually employ such type of inference for predicting intervals such as prediction intervals, tolerance

intervals and simultaneous tolerance intervals. The practitioner's method has two properties. First of all, the estimation does not take into account the sample size, so they must take into account the asymptotic notion of these intervals. The second property is that the conventional method explained in 5.1.1 is usually used for estimating a point-wise confidence interval for the conditional mean $f(x)$, and thus it cannot be used to estimate the conditional response variable $Y(x)$, which has also an error term. For a detailed discussion on these statistical intervals and their differences see [Hahn 91] and [Krishnamoorthy 09].

- **Inference on the conditional mean function** $f(x)$:
  - *Asymptotic point-wise inference:* this is the definition of the confidence interval for the mean regression function $I_{\beta}^{pw}(x)$ explained in 5.1.2 which contains, asymptotically, a desired proportion β of the conditional distribution of the estimated mean function $\hat{f}(x)$ for each combination of the predictors.
  - *Simultaneous Confidence based inference on the mean regression for all* $x \in \mathcal{X}$*:* this is the idea of γ-confidence band $I_{\gamma}^{cb}(x)$ for the regression function described in 5.2.2. These intervals create an envelope around the entire mean regression function $f(\cdot)$ such that, for all $x \in \mathcal{X}$, the probability that the true $f(x)$ is contained in the band is simultaneously γ.
- **Inference on the response variable** $Y(x) = f(x) + \varepsilon$:
  - *Asymptotic point-wise inference:* this is the definition of prediction interval for regression $I_{\beta}^{Pred}(x)$ explained in 5.2.1 which contains, asymptotically, a desired proportion β of the conditional distribution of the response variable $Y(x)$ for each combination of the predictors.
  - *Confidence based point-wise inference:* this is the definition of the tolerance interval for regression $I_{\gamma,\beta}^{T}(x)$ explained in 5.2.3. The interval contains, with a confidence level γ, at least a desired proportion β of the conditional distribution of the response variable $Y(x)$ for each combination of the predictors.
  - *Simultaneous confidence based inference on the response variable for all* $x \in \mathcal{X}$*:* this is the idea behind β-content γ-coverage simultaneous regression tolerance intervals $I_{\gamma,\beta}^{TS}(x)$ described in 5.2.4. These intervals create an envelope around the entire mean regression function $f(\cdot)$ such that, for all $x \in \mathcal{X}$, the probability that $Y(x)$ is contained in the band is β, and this coverage is guaranteed with a confidence level γ.

Note that tolerance intervals and simultaneous tolerance intervals for least squares regression have been well studied for linear regression but the application of these methods in the non-linear and particularly the non-parametric case are limited in the literature. Figure 5.8 displays these intervals for a simple linear model.

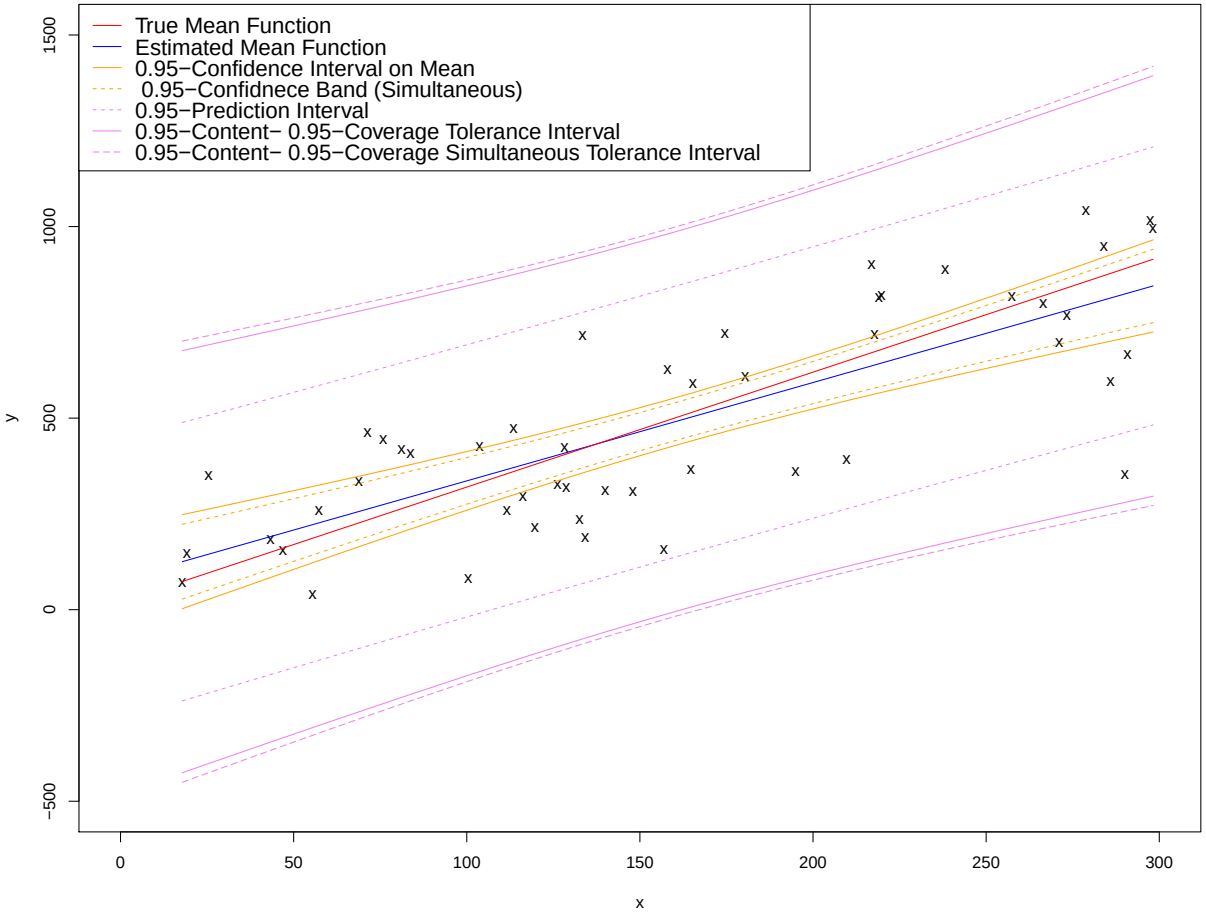

Figure 5.8: Comparing different interval prediction methods in linear least-sqaures regression.

5.4.2 Quantile Regression Models

A quantile regression model can estimate one conditional quantile so one-sided and two-sided interval estimation is treated separately.

- **One-sided interval prediction:**
  - *Estimates of point-wise interval:* this is the definition of quantile regression explained in 4.4.1 and these intervals are similar to one-sided prediction intervals in least-squares models defined in 5.2.1.
  - *Confidence based point-wise inference:* this is the definition of confidence intervals on regression quantiles explained in 5.3.1. The obtained one-sided

interval contains, with a confidence level γ, at least a desired proportion β of $Y(x)$. They have so far been studied for linear models, and they are are similar to one-sided tolerance intervals for regression explained in 5.2.3.

- **Two-sided interval prediction:** in order to obtain two-sided $(1-\alpha)$-content conditional intervals, one must build two distinct quantile regression models: a lower $\frac{\alpha}{2}$-quantile regression model and an upper $(1-\frac{\alpha}{2})$-quantile regression model.
  - *Estimates of point-wise interval:* This is done by a pair of upper and lower quantile regression model. These intervals are estimations and they are similar to two-sided prediction intervals in least-squares models defined in 5.2.1.
  - *Confidence based point-wise inference:* These two-sided intervals contain, with a γ confidence level, a proportion $1-\alpha$ of $Y(x)$. As noted, we need two quantile regression models but each model now itself needs a confidence interval, as explained in 5.3.3. There is not any work using this method, and they are are similar to the two-sided γ-coverage $(1-\alpha)$-content least-squares tolerance intervals explained in 5.2.3.

  Two-sided interval prediction is more meaningful by enforcing the *non-crossing* constraint but this may lead to a non-convergent quantile estimator. Thus we have to choose between "non-correct" or non-convergent estimators.

5.5 Conclusion

This chapter discussed several methods of finding intervals in regression models which contain a desired proportion of the response variable. One common category contains methods that consist of building a least squares or mean estimating regression model and then employing some kind of statistical inference technique to predict such intervals. Another classical method is based on quantile regression models. We reviewed different types of intervals and described their frequentist interpretation. We also take advantage of this chapter to address a common interval prediction method in the Machine Learning community. *Unfortunately, most practitioners of this community usually employ this conventional inference for predicting interval such as prediction intervals, tolerance intervals and simultaneous tolerance intervals. We dedicated the first section of this chapter to explaining this conventional technique and its applications, and to reviewing its drawbacks.*

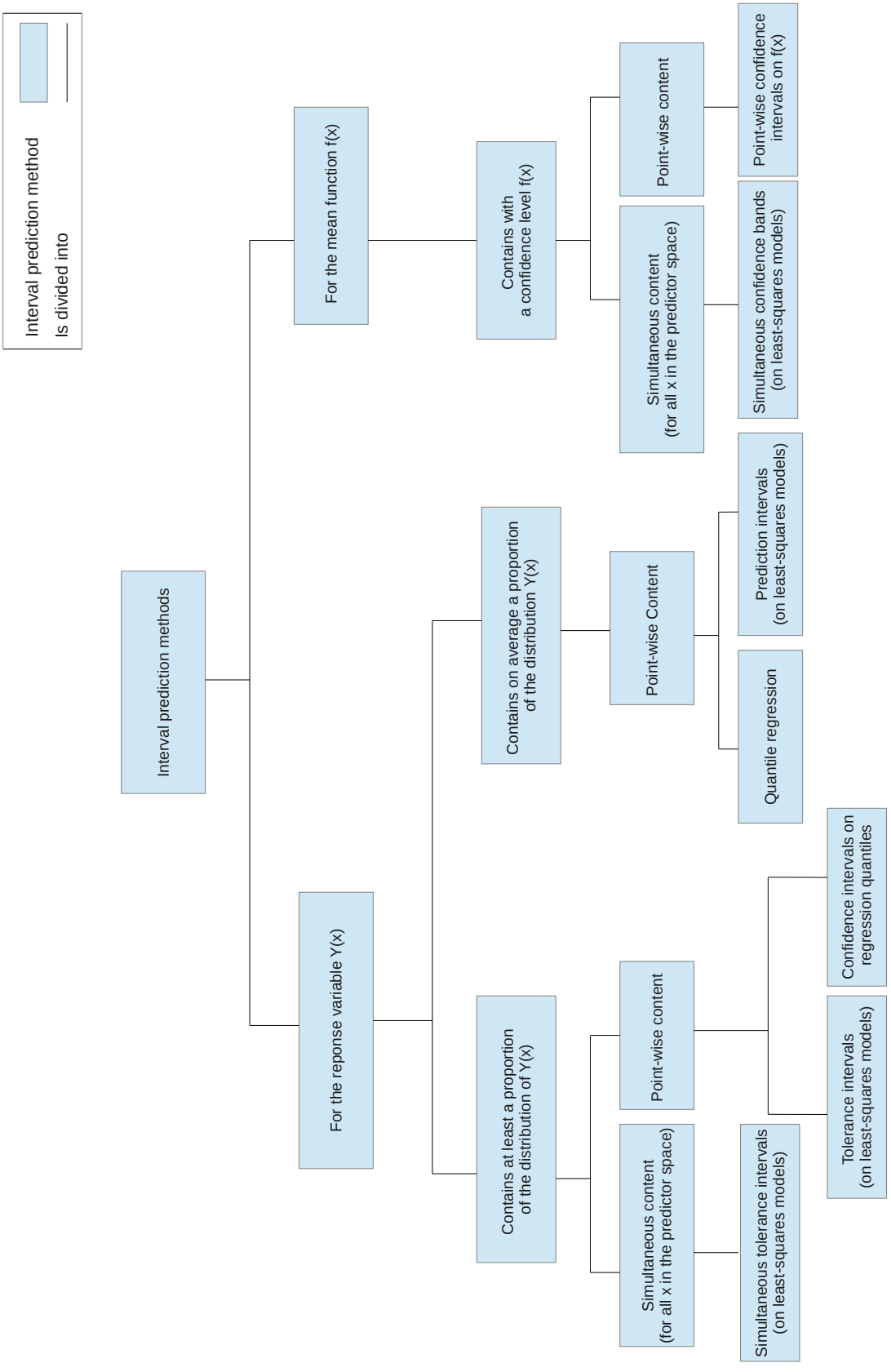

Figure 5.7: A classification of the statistical interval prediction methods in the regression context.

Chapter 6

Predictive Interval Framework

Contents

The previous chapter discussed different concepts and methods of interval prediction within the regression context and we referred to such methods with the "interval prediction method" term. The goal of this chapter is to propose a new interval prediction framework. We introduce the concept of regression predictive intervals and regression predictive interval models. Next, we propose a statistical test to verify if an "interval prediction model" is a "predictive interval model". In the same context, we introduce two measures for rating interval prediction models. These measures rate the efficiency and the tightness of the obtained envelope. Next, we describe the relationship of predictive intervals models and tolerance intervals for regression and confidence interval on quantile regression. We explain how to choose a confidence level γ to obtain efficient and reliable predictive intervals models.

The final part, is dedicated to an illustrative example which compares two distinct interval prediction methods on the motorcycle dataset [Silverman 85].

6.1 Interval Prediction Models

The goal of this paragraph is to emphasize the differences between intervals, interval prediction methods and interval prediction models. An interval prediction method is the procedure required for obtaining an interval. It is just the way to do it but when it is applied to a dataset, we obtain an interval prediction model. For example, a tolerance interval for regression is a type of interval. The method to obtain it in linear models is described in [Krishnamoorthy 09] and, when applied to a dataset, the model which gives the tolerance interval for each point in the predictor space is the interval prediction model.

Definition 19 *A regression β-content interval prediction model, built on the dataset $\mathcal{S}$, is function $I(\cdot)_{\mathcal{S},\beta}$ from the predictor space $\mathbb{R}^P$ to the response variable space $\mathbb{R}$ such that:*

$$I(\cdot)_{\mathcal{S},\beta} : \mathbb{R}^P \to \mathbb{I}, \text{ where } \mathbb{I} = \{[a,b] | a,b \in \mathbb{R} \cup \{-\infty, \infty\}, a < b\}. \tag{6.1}$$

and, the expected content of the intervals is at least β:

$$E_{\mathcal{S}}\left(P\left(Y(x) \in I(x)_{\mathcal{S},\beta} \middle| \mathcal{S} \right) \right) \geq \beta. \tag{6.2}$$

Thus when the *size of our training set goes to infinity* and under certain conditions, a β-content interval prediction model finds intervals that on average contain, at least, a β of the distribution of $Y(x)$. *This is a quite broad definition which covers all the interval prediction method for $Y(x)$ and we will use it for such purpose.*

Our works deals with the regression models, so we omit to mention the regression word and use "interval prediction model" instead of "regression interval prediction model". Note that test and model selection techniques are always applied to models and not to methods. However when a method is more efficient than its competitors on several datasets or in a general theoretical framework, we can state that this method is more efficient than others. Section 6.3 introduces a test for predictive interval models. The next section introduces predictive intervals and predictive interval models. Chapter 9 uses several regression datasets to compare different interval prediction methods. In Chapter 9, we will use different datasets to check whether the compared interval prediction methods are, in general, reliable enough to be used as predictive interval methods.

6.2 Predictive Interval Models

In the frequentist interpretation of confidence intervals, a confidence interval for a parameter contains zero or one parameter. The parameter is fixed and confidence intervals change with different random samples. In the same way, the γ used in tolerance intervals for regression

defined in (5.17) and confidence intervals for regression quantiles formulated by Equations (5.28) and (5.29) mean the following: the probability that the obtained intervals contain, under re-sampling, at least a proportion β of the conditional distribution of the response value $Y(x)$ is γ. We know that the confidence level in Neyman-Pearson confidence intervals is independent of the observed sample. It means that if we obtain γ-confidence β-content tolerance intervals of an observed sample from a regression function, then the confidence level γ does not induce any posterior probability of including β proportion of the distribution of $Y(x)$. Therefore, the confidence coefficient in frequentist confidence intervals cannot be interpreted as posterior probability. This idea is discussed in detail in Chapter 7 in [Walley 91].

Hence, under the frequentist viewpoint of regression, the true conditional response variable's inter-quantile is included with probability zero or one in the obtained interval (by using tolerance intervals for regression or confidence intervals for regression quantiles). Our goal is to obtain intervals that correctly bracket these inter-quantiles. They can be found in two ways: the first approach takes a very high confidence level like $\gamma \approx 1$ and the second method finds the smallest confidence level $0 < \gamma_0 < 1$ which includes the true unknown model. We introduce the concept of predictive intervals which refer to both of these intervals. A predictive interval built on $\mathcal{S}$, is guaranteed to contain for the query point x, at least a desired proportion of the conditional distribution of the response variable. *It can be obtained with tolerance intervals for regression or confidence intervals for regression quantiles but these concepts have so far only been treated for linear models*

Definition 20 *Let $\mathcal{S} = \{(x_1, Y_1) \cdots, (x_n, Y_n)\}$ denote a random sample where $Y_i = f(x_i) + \varepsilon_i$ and ε_i is white noise. A β-content predictive interval for x, denoted by $I(x)^P_\beta$, is an interval such that:*

$$P_{Y(x)}\left(Y(x) \in I(x)^P_\beta \middle| \mathcal{S}\right) \geq \beta, \text{ where } I(x)^P_\beta = [L(x)^P_\beta, U(x)^P_\beta]. \tag{6.3}$$

Since we have observed $\mathcal{S}$, $I(x)^P_\beta$ is no longer random and the probability measure is just related to cover at least a proportion β of the conditional distribution of the response variable $Y(x)$ for a specified combination of the predictors.

Definition 21 *Let $\mathcal{S} = \{(x_1, Y_1) \cdots, (x_n, Y_n)\}$ denote a random sample where $Y_i = f(x_i) + \varepsilon_i$ and ε_i is white noise. A β-content predictive interval model, denoted by $I(\cdot)^P_\beta$, is a function such that:*

$$I(\cdot)^P_\beta : \mathbb{R}^p \to \mathcal{I}, \text{ where } \mathcal{I} = \{[a, b] | a, b \in \mathbb{R} \cup \{-\infty, \infty\}, a < b\}, \tag{6.4}$$

and for all x in the predictor space, the obtained interval is a β-content predictive interval described by (6.3).

6.3 Predictive Model Test

The goal of this part is to develop an statistical test with which we can rate the reliability of any model claiming to provide β-content predictive intervals. A predictive interval model must provide predictive intervals for each point in the predictor space. We saw that the distribution of $Y(x)$ changes for each value of x. So in order to see whether an interval for the regression function at the point x contains at least a proportion β of the s distribution of $Y(x)$, we need (for each combination of predictors x) a sample set from the distribution of $Y(x) = f(x) + \epsilon$, and then we can observe if the constructed interval contains a proportion s distribution of the distribution of $Y(x)$. In the same way, in order to verify if the methods works for an entire dataset $\{x_i| \in (1, \cdots, n)\}$, we need a distinct sample set for each x_i and this sample must be drawn from $Y(x_i)$. Since a sample set is required for each $x_i, i \in (1, \cdots, n)$, the described procedure requires a huge dataset having many observations for each point x in the feature space which makes it impractical or impossible for multiple regression problems. However, we can make some approximations and use the results stated above to derive the following test. We first begin by defining a variable MIP on the dataset. We will see that this variable can be approximated by a normal distribution. Then we use the normal distribution to define the α level predictive model test. So one can verify for example if tolerance intervals for the linear regression applied to the Motorcycle dataset is a predictive interval model or not.

6.3.1 Simultaneous Inclusion with Predictive Intervals

A β-content predictive interval for regression $I(x)^P_\beta$ must contain at least β proportion of the conditional distribution $Y(x)$. Hence, the probability measure in (6.3), is just related to contain at least a proportion β of the conditional distribution of the conditional response variable $Y(x)$. We define the function $V(x)$ as:

$$V(x) = \begin{cases} 1 & \text{if } Y(x) \in I(x)^P_\beta, \\ 0 & \text{otherwise.} \end{cases}$$

The above definition means that the probability that $V(x)$ is equal to 1 is β, so $V(x)$ has a Bernoulli distribution with $p = \beta$.

$$V(x) \sim Bernoulli(\beta). \tag{6.5}$$

Suppose that we have a dataset of n_{train} observations $\mathcal{T} = \{(x_1, Y_1) \cdots, (x_{n_{train}}, Y_{n_{train}})\}$ with which we build our model, and n_v other observations $\mathcal{S} = \{(x^v_1, Y^v_1) \cdots, (x^v_{n_v}, Y^v_{n_v})\}$, not contained in the original dataset as our test set. If we apply the function $V(\cdot)$ on the whole test set $\mathcal{S}$ and sum the result, we obtain:

$$MIP_{\mathcal{S},\beta} = n_v^{-1} \sum_{i=1}^{n_v} V(x^v_i). \tag{6.6}$$

Therefore, we can deduce that $MIP_{\mathcal{S},\beta}$ has a Binomial distribution. This is expressed formally by (6.7) where $\mathcal{B}(n_v, \beta)$ is a binomial distribution with $n = n_v$ and $p = \beta$.

$$n_v MIP_{\mathcal{S},\beta} \sim \mathcal{B}(n_v, \beta). \tag{6.7}$$

If n_v is sufficiently large, we can assume that $MIP_{\gamma,\beta}$ has a normal distribution as:

$$MIP_{\mathcal{S},\beta} \sim \mathcal{N}(\beta, \frac{\beta(1-\beta)}{n_v}). \tag{6.8}$$

Thus the fraction of instances having their response value included in their predictive intervals is on average β. This means such predictive intervals for regression have in average a simultaneous content of β so, on average, they are like simultaneous regression tolerance intervals. **For small to medium datasets, $MIP_{\mathcal{S},\beta}$ is computed in a cross-validation or leave-one-out schema on the whole dataset which means that $\mathcal{S} = \mathcal{T}$.**

6.3.2 Testing Predictive Interval Models

As we have seen in (6.8), the random variable $MIP_{\mathcal{S},\beta}$ can usually be well approximated by a normal distribution. The test below is used to verify, with level α, if the interval prediction method does provide β_0-content predictive intervals for all x in the predictor space.

$$H_0 : \beta \geq \beta_0 \ \textit{versus}\ H_1 : \beta < \beta_0, \tag{6.9}$$

then H_0 can be rejected with significance α where:

$$\textbf{MIP Test: } MIP_{\mathcal{S},\beta} < n_v^{-1/2}\beta_0(1-\beta_0)Z_\alpha = F^{\alpha}_{\beta_0,n_v}, \tag{6.10}$$

where Z_α is the α-quantile of the standard normal distribution. So if (6.10) is true, we fail to reject the null hypothesis with significance level α, and accept the underlying model as a model providing β-content predictive intervals.

In this thesis we used a significance level of $\alpha = 0.05$, so for each dataset we compared the $MIP_{\mathcal{S},\beta}$'s value on the test set with $F^{0.05}_{\beta_0,n_v}$. For the sake of simplicity, we refer to $MIP_{\mathcal{S},\beta}$ for a given dataset $\mathcal{S}$ and desired proportion β as MIP (Mean Inclusion Percentage). As we have seen in (6.8), the fraction of response values inside their β-content predictive intervals converges to β, so the test defined in (6.10), where $\alpha = 0.05$ is used to verify if the obtained intervals, with a confidence level 0.95 *and on average and not at least like in simultaneous tolerance internals*, do simultaneously contain a proportion β_0 of the distribution of $Y(x)$ for all x in the predictor space.

6.4 Comparing Interval Prediction Models

The above test can be used to verify the reliability of a model claiming to provide β-predictive intervals but it does not tell us anything about its efficiency. For a given dataset,

we may have several interval construction models which pass this test but we need to find the the most efficient one. For this purpose, we define the dataset measures listed below. These measures are then used as building blocks for some graphical charts and plots explained further in this section. The idea is to provide graphical tools which can help us to compare the effectiveness of different interval prediction methods through different datasets. Each symbol denotes a variable, that when applied to an interval prediction model, is indexed by the predictive interval model's dataset and method. For example MIS denotes the Mean of Interval Size. However $MIS_{\mathcal{S}}^{m}$, denotes the mean of intervals size of intervals obtained with the interval prediction method m applied to the dataset $\mathcal{S}$.

6.4.1 Direct Dataset Measures

For each of the datasets the following quality measures can be computed:

- MIP: Mean Inclusion Percentages and must satisfy the MIP constraint:

$$\textbf{MIP Constarint: } MIP_{\mathcal{S},\beta} \geq F_{\beta,n}^{0.05}.$$

  (see (6.6)).

- MIS: Mean of Interval Size.

$$MIS = \frac{1}{n}\sum_{i=1}^{n} size(I(x)_{\beta}^{P}).$$

- σ_{is}: sample standard deviation of interval sizes.

$$\sigma_{is} = \frac{1}{n}\sum_{i=1}^{n}(size(I(x)_{\beta}^{P}) - MIS)^2,$$

where $size(I(x)_{\beta}^{P})$ refers to the size of the β-content predictive interval. For small to medium datasets, the above measures are computed using a cross-validation or a leave-one-out schema.

6.4.2 Composed Dataset Measures

We use the above quality measures to define the following composed measures:

Normalized MIS

Suppose that we want to test c different methods ("$Method_1$", "$Method_2$" ,..., "$Method_c$") on the dataset $\mathcal{S}$. They give us c distinct models and each model has a Mean of Interval Size (MIS), so we have: $MIS_{\mathcal{S}}^{1}$, $MIS_{\mathcal{S}}^{2}$,..., $MIS_{\mathcal{S}}^{c}$. But depending on the dataset and β's value, one model may satisfy the MIP constraint or not. For a model that does not pass the test, its normalized MIS value is not computed. For each model satisfying its constraint

on the dataset, its normalized MIS is equal to the ratio of its MIS to the maximum MIS on this dataset

$$normalizedMIS^i_{\mathcal{S}} = \frac{MIS^i_{\mathcal{S}}}{\underset{i\in(1,\cdots,c)}{max}(MIS^i_{\mathcal{S}})}.$$

If we have :

$$MIS^{m_1}_{\mathcal{S}} \geq MIS^{m_2}_{\mathcal{S}} \text{ and } MIP^{m_1}_{\mathcal{S},\beta} \geq F^{0.05}_{\beta,n} \text{ and } MIP^{m_2}_{\mathcal{S},\beta} \geq F^{0.05}_{\beta,n}$$
$$\Leftrightarrow m_1 \textbf{ provides a wider reliable envelope than } m_2\textbf{.}$$

M_2 is better than m_1 because it satisfies the MIP constraint and it also gives the smallest normalized MIS value. Choosing the ratio to the maximum MIS value rescales the MIS value between 0 and 1 and lets us compare the strength of methods across different datasets. However we can not use the normalized MIS to compare two models (constructed on the same dataset) that obtain different MIP values but have equal or approximately equal MIS values. In this case, we have to compare them by their Equivalent Gaussian Standard Deviation, explained below.

Equivalent Gaussian Standard Deviation (EGSD)

If we have two reliable models (constructed on the same dataset) having different MIP values but approximately equal MIS values, we normally choose the one that gives the to higher MIP. But the situation can get more complicated for models (constructed on the same dataset) with different MIS values and different MIP values. EGSD is a measure which can be used to compare interval prediction models, constructed on the same dataset, which have different MIP values. Such models can have different or equal MIS values. Let m be a β-content interval prediction model built on the dataset $\mathcal{S}$, yielding $MIS^m_{\mathcal{S}}$ and $MIP^m_{\mathcal{S},\beta}$. The idea behind EGSD is to find the Equivalent Gaussian Distribution (EGD) for successful predicted intervals of m. We have seen that by taking intervals size on average equal to $MIS^m_{\mathcal{S}}$, that $MIP^m_{\mathcal{S},\beta}$ of the observations will be contained in their prediction interval. So EGD is the distribution of the size of predicted intervals obtained by model m_1 that correctly contains their response variable. Therefore the EGD which has the smallest variance corresponds to the most efficient model. The Equivalent Gaussian Distribution for m is the normal distribution θ-content inter-quantile size of which will be equal to $MIS^m_{\mathcal{S}}$. We have: $\theta = MIP^m_{\mathcal{S},\beta}$. So the Equivalent Gaussian Standard Deviation of m is calculated by:

$$EGSD^m_{\mathcal{S}} = \frac{MIS^m_{\mathcal{S}}}{2Z_{1-\frac{\alpha}{2}}\theta}, \text{ where } \theta = MIP^m_{\mathcal{S},\beta}\ , \alpha = 1-\beta,$$

and Z_α is the α-quantile of the standard normal distribution. Now by using each model's EGSD, we can compare models with different values of MIP and MIS. EGSD measures the trade-off between average interval size and the fraction of successful predictions. **Smaller EGSD values denote more effective interval prediction models.** Finally, for the sake of readability, all found EGSD are normalized in each dataset. Thus the final value

is the ratio of the method's $EGSD^m_{\mathcal{S}}$ to the maximum $EGSD$ value on the underlying dataset:

$$normalizedEGSD^m_{\mathcal{S}} = \frac{EGSD^m_{\mathcal{S}}}{\underset{i\in(1,\cdots,c)}{max}(EGSD^i_{\mathcal{S}})}.$$

Note that if the model m_1 has smaller EGSD than the model m_2, it does not mean m_2's envelope is wider than m_1's envelope. As seen above smaller normalized MIS values mean smaller envelopes and smaller EGSD values means more effective models.

6.4.3 Figures

Plots and charts help us to compare different interval prediction methods on different datasets because a figure can visualize complex and big tables. Each plot is dedicated to one dataset and it compares dataset measures of different interval prediction methods on the same dataset whereas a chart compares a desired dataset measure for different methods and across different datasets. All the presented plots have the same x axis. This axis is labeled "Nominal MIP", and it represents distinct values of the desired proportion (distinct β values). On the other hand, each plot type has a different y axis. This axis denote the underlying dataset measure on the tested interval prediction models.

MIP plot

The MIP plot is similar to the well-known Q-Q plot with the difference that it compares MIP instead of quantiles. The x axis is denoted by "Nominal MIP" and it represents the desired proportion of inclusion (distinct values of β). The y axis is denoted by "Obtained MIP". It represents the model's MIP. Each point represents the model's obtained MIP for its value on the "Nominal MIP" axis. This figure always has two lines: the "MIP constraint line" and the "Nominal MIP line". The "MIP constraint line" displays $F^{0.05}_{\beta,n}$ for different values of nominal MIP, and the "Nominal MIP line" represents the function $y = x$. By looking at this figure we can see the reliability of a method for different nominal MIP. *The first value in the x axis where a line crosses the MIP constraint line will be called its* ***failure MIP****. It is obvious that the method having the higher failure MIP is the most reliable one.*

One can also use the MIP plot to rate the model's **precision**. If a model obtains MIP values much higher than the desired nominal MIP, it means that the method is reliable but not precise. For example a model which obtains MIP values of 0.45, 0.9 and 0.99 for respective nominal MIP of 0.25, 0.75 and 0.95 is reliable but not precise. The most precise model is the one having the nearest line to the "Nominal MIP line". Finally, the best model in this plot is the one which is the most precise and the most reliable. It means that **the best model in a MIP plot is the one having the nearest line to the upper side of the "Nominal MIP line".** Figure 6.2 is an example of an MIP plot.

EGSD plot

EGSD plot: the y axis of an EGSD plot is labeled by "Normalized EGSD Value" and it represents the model's normalized EGSD value. By looking at this figure we can compare the efficiency of different models. It is obvious that the model having the highest line is the most inefficient model. *We suggest using this plot along with the MIP plot to rate the efficiency of reliable methods.* However one may ignore the reliability aspect and take advantage of this plot to compare the efficiency of different models.

MIS plot

The y axis of an EGSD plot labeled "Normalized MIS Value" and it represents the model's normalized MIS value. By looking at this figure, we can compare the model which obtains the tightest reliable envelope. The model having the highest line provides the widest envelope. If a model does not pass the MIP test, its normalized MIS value is not computed. The MIS plot shows each model's normalized MIS until its "failure MIP". *We suggest using this plot along with the EGSD plot.*

Charts

Charts are used to compare one dataset measure on different datasets. We propose the following charts:

- Mean Inclusion Percentage Chart (MIP chart): the goal of this chart is to compare the mentioned methods based on their fraction of response values located inside their predictive intervals. It just displays the MIP value and it usually does not contain important information.
- MIS ratio chart: this chart displays the normalized MIS measure on different datasets.
- Equivalent Gaussian Standard Deviation chart (EGSD chart): it displays the normalized EGSD measure on different datasets.

6.5 Predictive interval models with tolerance intervals and confidence interval on quantile regression

This section explains the relationship of predictive intervals models and tolerance intervals for regression and confidence intervals on quantile regression. The first part describes how the confidence level γ in tolerance intervals for regression and confidence intervals on quantile regression can be used to obtain predictive interval models. Then the second part will help us find the "best" value of γ. By this, we mean the confidence level that provides the predictive interval model with the smallest MIS value.

6.5.1 Simultaneous Inclusion

As seen previously, the confidence level γ in regression tolerance intervals defined in (5.17) is related to the estimated regression model $\hat{\mathcal{M}} = (\hat{f}, \hat{\sigma})$ and β is related to $Y(x)$. We know that the true regression model $\mathcal{M} = (f, \sigma)$ is unknown and the estimated model $\hat{\mathcal{M}} = (\hat{f}, \hat{\sigma})$ is a random vector which depends on the random sample $\mathcal{S} = \{(x_1, Y_1) \cdots, (x_n, Y_n)\}$. Let $\mathcal{R}_{M,\gamma}$ denote the γ-confidence region of $\mathcal{M}$. Then we have:

$$P_{\mathcal{M}}(\mathcal{M} \in \mathcal{R}_{M,\gamma}) = \gamma,$$

so $\mathcal{R}_{M,\gamma}$ is a subspace in the regression model space of $\mathcal{S}$. It contains regression models for which the probability of $Y(x)$'s β-inter-quantile being contained in the tolerance interval $I(x)^T_{\gamma,\beta}$ is γ. However, once $\mathcal{S}$ has been observed, $\hat{\mathcal{M}} = (\hat{f}, \hat{\sigma})$ and $I(x)^T_{\gamma,\beta}$ becomes non-random and there exists a value of γ_0 such that for all γ greater or equal to γ_0 the true model $\mathcal{M} = (f, \sigma)$ is included in $\mathcal{R}_{M,\gamma}$. It can be found by its γ_0 quantile and it is stated formally by:

$$P^{-1}_{\mathcal{M}}(\gamma_0) = \mathcal{M}.$$

Consequently, for a given sample set $\mathcal{S}$, we suppose that we have found a confidence level $\gamma \geq \gamma_0$ such that $\mathcal{M} \in \mathcal{R}_{M,\gamma}$. This allows us to ignore the external probability in (5.17) and we can state that, for the fixed value of covariate x, the probability of $Y(x)$ being in $I(x)^T_{\gamma_0,\beta}$ is greater or equal to β. Since we have observed $\mathcal{S}$, $I(x)^T_{\gamma,\beta}$ is no longer random and the resulting probability measure in (5.17), is just related to cover at least a proportion β of the conditional distribution of the response variable $Y(x)$ for a specified combination of the predictors. Then we have:

$$MIP_{\mathcal{S},\beta} \equiv MIP_{\mathcal{S},\gamma,\beta}.$$

If we have the right value for γ as stated before, the fraction of instances having their response value included on their regression tolerance intervals is in average β. *The above discussion holds similarly for γ confidence intervals for β-regression quantiles.* **Once we have found a confidence level γ greater than or equal to the true and unknown γ_0 stated above, the resulted γ-coverage β-content tolerance intervals for least squares regression (or γ-confidence β-content regression quantiles in the case of a quantile regression model) can be used as β-content predictive intervals.** Now, we can use the test defined above to find a γ greater or equal to γ_0. Note that the most reliable method is to take $\gamma \approx 1$ but it provides wide intervals, so our test is an intuitive way of finding the smallest reliable confidence level. As we have seen in (6.8), the random variable $MIP_{\mathcal{S},\gamma,\beta}$ can usually be well approximated by a normal distribution. For the sake of simplicity we refer to $MIP_{\mathcal{S},\gamma,\beta}$ for a given dataset $\mathcal{S}$, fixed values of γ and β as MIP (Mean Inclusion Percentage).

6.5.2 Hyper-parameter Tuning and Model Selection

This part addresses hyper-parameter tuning questions related to predictive interval explained in Section 6.2. We suppose that the β-content predictive intervals are obtained via tolerance intervals for regression or via confidence interval on regression quantiles. For the sake of brevity we continue this part with tolerance intervals for regression but the same procedure and statements hold for confidence interval on regression quantiles.

Tolerance intervals are obtained upon regression models which may themselves have hyper-parameters. Consider an example of constructing predictive interval models by tolerance intervals on a KNN regression. This model has two hyper-parameters, the KNN regression's hyper-parameter which is the number K, and the confidence level γ related to tolerance intervals. First we find the regression's hyper-parameters; It can be K for KNN or Loess, or kernel related parameters in SVM regression or nothing in the linear regression (this hyper-parameter depends on the regression method). Once we have found the best regression model, we use an iterative algorithm that searches the smallest γ that satisfies the tuning constraint defined below which results in intervals having the smallest MIS measure. MIP and MIS are computed based on a leave-one-out or 10-fold cross validation scheme on the training set.

$$\textbf{Tuning Constraint: } MIP_{\mathcal{S},\gamma,\beta} = \beta. \tag{6.11}$$

A more conservative approach could be to find γ which gives the tolerance intervals having the smallest mean interval size and also satisfying the constraint below where $Z_{1-\alpha}$ is the $(1-\alpha)$-quantile, i.e. 0.95 quantile, of the standard normal distribution.

$$\textbf{Conservative Tuning Constraint: } MIP_{\mathcal{S},\gamma,\beta} \geq t^{-1/2}\beta(1-\beta)Z_{1-\alpha}. \tag{6.12}$$

High values of γ will guarantee the tuning constraint (6.11) but the computed intervals can be very large, so the search begins with a high confidence value like $\gamma = 0.9$ or $\gamma = 0.99$ and we try to decrease γ and thus decrease the mean interval size. This procedure is repeated as long as the tuning constraints are satisfied and the search strategy is left to the user. **Note that the tuning constraint is a hard constraint and there is no trade-off between satisfying this constraint and minimizing the MIS.** Some datasets might require just 2 or 3 iterations but some others may work with small γ. It depends on the dataset and it can be influenced by the domain expert.

6.6 Illustration

Figure 6.1 gives an illustration of two distinct models used to obtain two-sided 0.95-content predictive intervals. These models are built with a 10-fold cross validation schema on the motorcycle dataset [Silverman 85]. The first one is a β-content predictive interval model which is constructed with confidence intervals on quantile regression [Kocherginsky 05]. The 0.95-level two-sided content is obtained with two different quantile regression models

as explained in 5.3.3. The second model is based on the conventional interval prediction method explained in 5.1.1 obtained with a least-squares SVM regression. The results of this experiment are displayed in Table 6.1. We can see that both models have their MIP value greater than $F^{0.05}_{0.95,133} = 91.89$, so they satisfy (6.10) and they are two-sided 0.95-content predictive interval models.

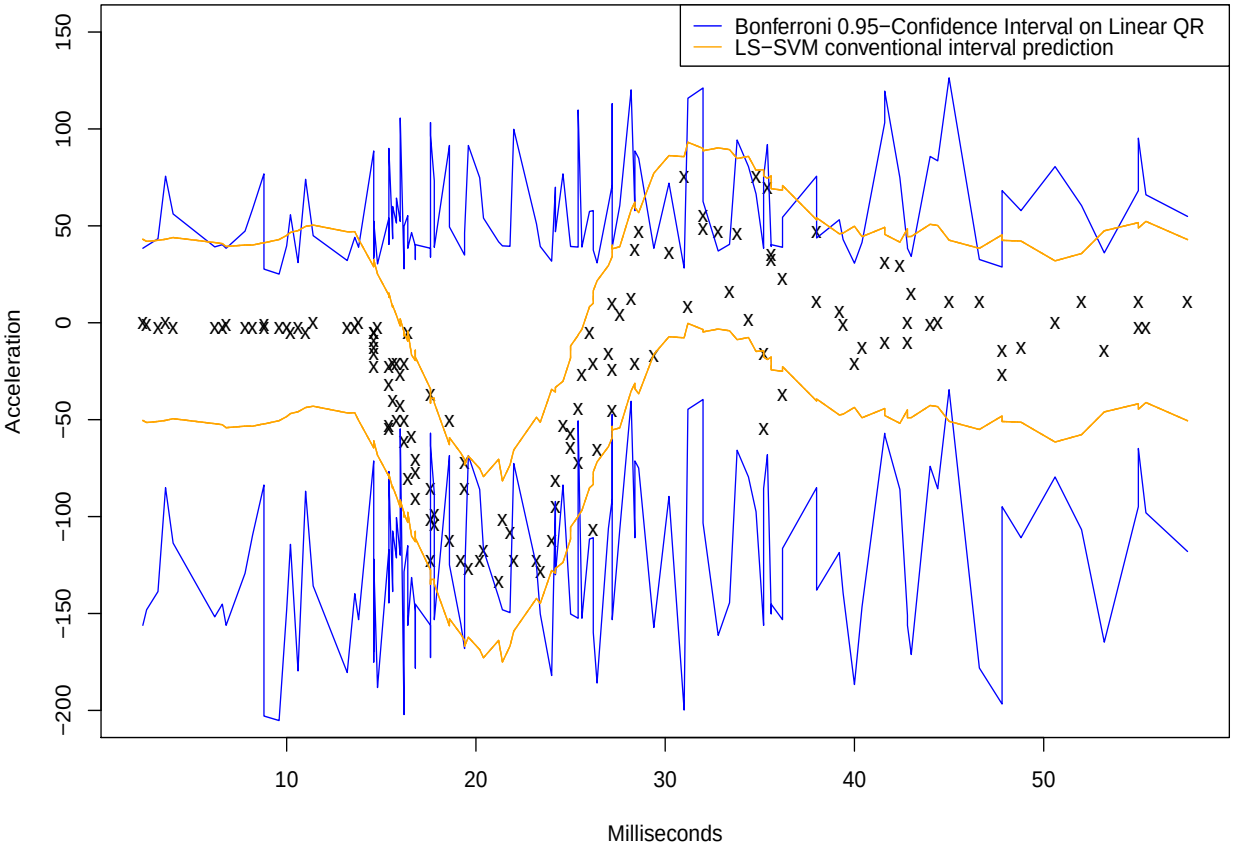

Figure 6.1: Two-sided 0.95-predictive intervals for the motorcycle dataset [Silverman 85].

However, their normalized MIS values show that the conventional model obtains much a tighter envelope than its competitor. Then the normalized EGSD value shows that the intervals obtained by the conventional model are also much more efficient than the intervals obtained by the two-sided Bonferroni model. Figure 6.1 represents the MIP plot for these

Method	MIP	MIS	σ_{is}	normalized MIS	EGSD
Two-sided Bonferroni method	92.46	182.05	5.13	1	1
LS-SVM conventional method	94.77	93.46	0	0.51	0.47

Table 6.1: Experiment results of Figure 6.1.

methods. We can see that the conventional method is very reliable on the motorcycle dataset. Although this method works here, we do not recommend it for this purpose. In Chapter 9 we will use several distinct datasets to demonstrate that the conventional method does not provide reliable models and must therefore not be used as predictive interval models.

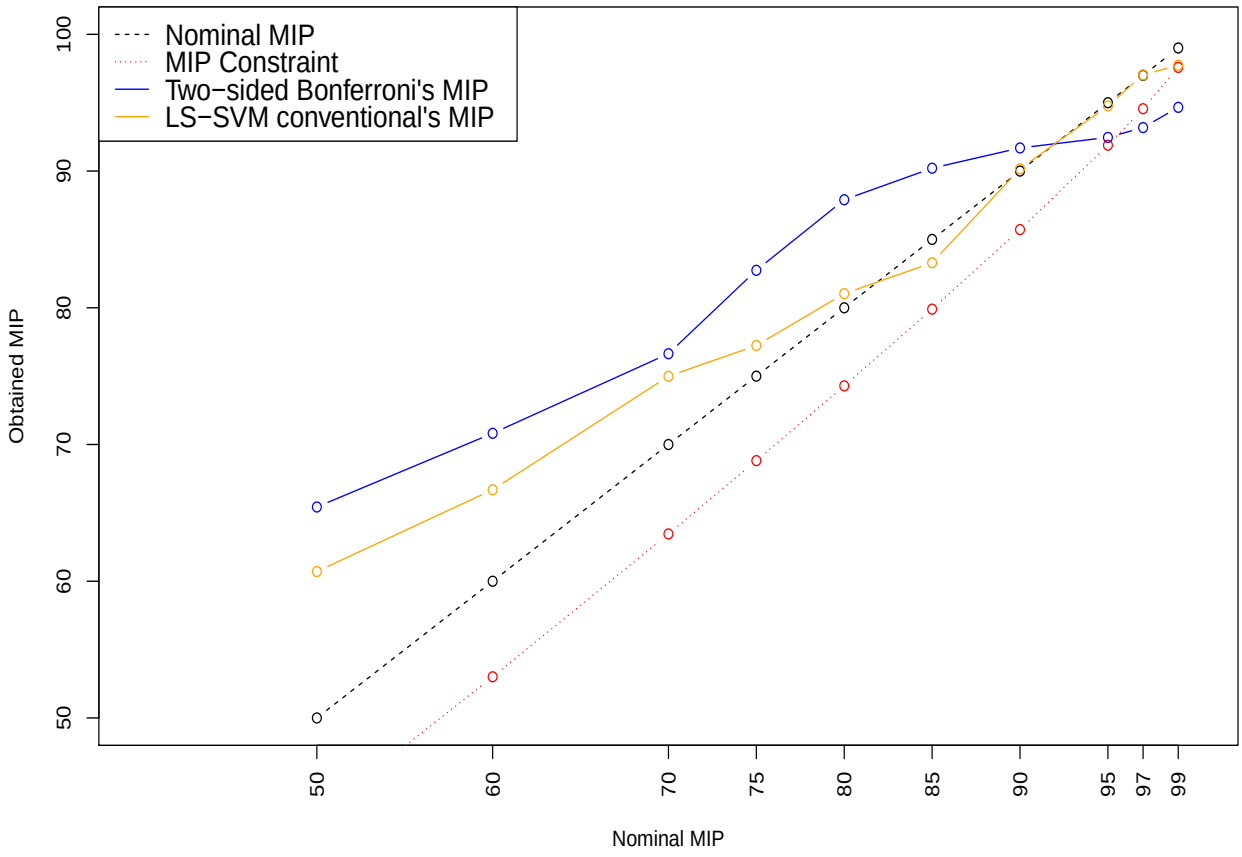

Figure 6.2: Comparing obtained MIP to the MIP constraint for different β values.

6.7 Conclusion

This chapter proposed a new interval prediction framework. We introduced the following notions: predictive interval's concept, predictive interval model's notion, a predictive interval model test and two interval prediction measures. We have seen that a predictive interval for linear models can be obtained with tolerance intervals for regression and confidence intervals on quantile regression. However, such models may provide wide intervals. So we explained how to tune the confidence level of tolerance intervals for regression and confidence intervals on quantile regression in order to obtain efficient and reliable predictive interval models. The final part gives an illustrative example which compares two distinct interval prediction methods on the motorcycle dataset [Silverman 85]. In the next chapter we take advantage of the test defined here to propose some non-parametric predictive interval methods. Then our efficiency measures will be used to demonstrate the superiority of the suggested non-parametric methods compared to their competitors.

Chapter 7

Predictive Interval Models for Non-parametric Regression

Contents

In the previous chapter we introduced the predictive interval framework. The contributions of this chapter are the introduction of two predictive interval methods for non-parametric regression. They are applied for two-sided interval prediction but one can also use them in a one-sided interval prediction context. We propose two predictive interval models for local linear regression which both give variable size intervals. We assume that the mean regression function is locally linear and the prediction error is locally homoscedastic. Our methods do not neglect the regression bias and find intervals that work properly with biased regression models. The proposed predictive intervals are constructed based on the leave-one-out or 10-fold cross validation prediction errors of the local linear regression. The local linear regression needs a regression bandwidth which could be found by any of the existing methods in the literature. In order to obtain our non-parametric predictive

intervals, we need a second bandwidth, which is the tolerance interval bandwidth (LHNPE bandwidth). This work suggests two different tolerance interval bandwidths: a bandwidth having a fixed number of neighbors and a bandwidth having a variable one but both obtain variable size intervals. In the end, we will see that all these methods can also be used for possibilistic regression with crisp input and output data. This chapter is organized as follows: the first section explains how to compute tolerance intervals for local linear regression. The next section describes how to use the tolerance intervals to obtain predictive interval models. Then we will briefly see how to obtain our interval prediction models with a commonly-used local linear regression method called Loess. Finally we will have an illustration section and we conclude this chapter with a comparison of our method to existing ones.

7.1 Tolerance Interval for local linear regression

We have seen in the previous chapter that predictive intervals can be obtained with tolerance intervals for regression or confidence intervals for regression quantiles. However these intervals have not yet been studied for non-parametric models, so in this section we introduce two new methods for the calculation of predictive intervals for local linear regression. Another important subject is the regression bias. It is well known that the optimal smoothing in LLR or other non-parametric regression methods consists of a trade-off between the bias and the standard deviation. This non-parametric bias does not vanish even with large sample sizes, so it is important to use methods that do not ignore the regression bias. The idea behind our predictive intervals is to exploit the local density of prediction error $(Y_i - \hat{f}(x_i))$ in the LHNPE neighborhood of the query point x^* to find the most appropriate interval that contains the desired proportion of response values $Y(x^*)$. The response variable predictive intervals are constructed by adding the regression estimates to the locally approximated prediction error's predictive interval. Prediction error's predictive interval are centered on negative bias, so when added to the biased regression results, they remove the regression bias. Thus it leads to response variable's predictive intervals which correctly contain a proportion β of the distribution of $Y(x)$.

7.1.1 Theoretical context

This part describes the theoretical context of tolerance intervals for local linear regression. We first define the concept of a Local Homoscedastic Normal Prediction Error (LHNPE) regression estimator. Then, we define the LHNPE neighborhood of a query point and in the end we will use a simple a straightforward inference to obtain the formula of tolerance intervals for local linear regression.

Definition 22 *The oscillation of the function $f : X \to \mathbb{R}$ on an open set U is defined as:*

$$\omega_f(U) = \sup_{x \in U} f(x) - \inf_{x \in U} f(x).$$

Definition 23 *A regression estimator $\hat{f}(x)$ is a **Local Homoscedastic Normal Prediction Error (LHNPE)** if it satisfies the following conditions:*

- *Normal distribution: the prediction error $\varepsilon_x^{pred} = Y(x) - \hat{f}(x)$ has a normal distribution.*
- *Almost constant distribution the prediction error: We suppose that the mean $\mu(\varepsilon_x^{pred})$ and the standard deviation $\sigma(\varepsilon_x^{pred})$ of the distribution for the prediction error have small local oscillations. This is defined formally as:*

  For all x, there exists an open set $U \ni x$, such that:

  $$\omega_{\mu(\varepsilon_x^{pred})}(U) \leq \upsilon_1 \text{ and } \omega_{\sigma(\varepsilon_x^{pred})}(U) \leq \upsilon_2,$$

  where υ_1 and υ_2 are small fixed positive values.

Definition 24 *Let $\hat{f}(x^*)$ be a LHNPE regression estimator for the query point x^*. The **LHNPE neighborhood for** x^* are instances for which the prediction error satisfies the LHNPE conditions. This neighborhood is described as below:*

$$Kset_{x^*} = \{(x_i, Y_i) | d(x^*, x_i) \leq b\}, \tag{7.1}$$

where $d(x^*, x_i)$ is a distance function in the feature space and b denotes the LHNPE bandwidth.

Note that the LHNPE bandwidth $Kset_{x^*}$ is different from the regression bandwidth Reg_{x^*} in local linear regression:

$$Reg_{x^*} = \{(x_i, Y_i) | d(x^*, x_i) \leq b_{reg}\}. \tag{7.2}$$

The regression bandwidth minimizes the regression bias-variance trade-off but the LHNPE bandwidth is used to find the neighborhood which satisfies the LHNPE conditions. The LHNPE neighborhood is almost always included in the regression neighborhood:

$$Kset_{x^*} \subseteq Reg_{x^*}, \tag{7.3}$$

because the constant's $\left(Y(x^*) - \hat{f}(x^*)\right)$ distribution in the neighborhood of the query point x^* usually occurs inside its regression neighborhood. It is possible to find two different regression neighborhoods being next to each other having approximately the same prediction error distribution and not the same regression neighborhood. There are already several references on regression bandwidth Reg_{x^*} selection in non-parametric regression. We do not treat this problem and the reader can find more details in [Fan 96] and [Härdle 90].

Proposition 10 *Let $Y(x) = f(x) + \varepsilon_x$ denote a regression function and let $\hat{f}(x)$ denote its Local Linear regression estimator. If our regression estimator satisfies the conditions below:*

- *Normal error distribution: $\varepsilon_x \sim \mathcal{N}(0, \sigma_x^2)$.*

- *Normal distribution of the local linear estimator:* $\hat{f}(x) \sim \mathcal{N}\Big(f(x) + Bias_{\hat{f}(x)}, \sigma^2_{\hat{f}(x)}\Big)$, *Fan et al. [Fan 95] have shown that this assumption holds under certain regularity conditions.*
- $\hat{f}(x)$ *satisfies the LHNPE conditions defined above.*

where $Bias_{\hat{f}(x^*)} = E[\hat{f}(x^*) - f(x^*)]$ *is the estimator's bias,* σ^2_x *is the variance of the error and* $\sigma^2_{\hat{f}(x^*)}$ *is the variance of the estimator. Then the* γ*-confidence* β*-content regression tolerance interval for the query point* x^* *is:*

$$I(x^*)^T_{\gamma,\beta} = \hat{f}(x^*) + I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}, \tag{7.4}$$

$$\text{where } \varepsilon^{pred}_{x^*} = Y(x^*) - \hat{f}(x^*).$$

In the above equation, $I(x^*)^T_{\gamma,\beta}$ and $I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}$ denote, respectively, the regression tolerance interval and the prediction error tolerance interval.

Proof: The assumptions lead to assume that the prediction error has a normal distribution and its variance is approximately the same in the neighborhood of x^*. Let x^* denote the query point and let $\varepsilon^{pred}_{x^*}$ denote its prediction error, then we have:

$$\varepsilon^{pred}_{x^*} = \varepsilon + f(x^*) - \hat{f}(x^*),$$

which results in:

$$\varepsilon^{pred}_{x^*} \sim \mathcal{N}(-Bias_{\hat{f}(x^*)}, \sigma^2_{x^*} + \sigma^2_{\hat{f}(x^*)}). \tag{7.5}$$

The tolerance interval of the prediction error, is denoted by

$$I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta} = [L(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}, U(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}] = ICentered(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta} - \widehat{bias}_{\hat{f}(x^*)},$$

$$\widehat{bias}_{\hat{f}(x^*)} = \frac{1}{card(Kset_{x^*})} \sum_{x_i \in Kset_{x^*}} \varepsilon^{pred}_i,$$

where $ICentered(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}$, $card(Kset_{x^*})$ and $\widehat{bias}_{\hat{f}(x^*)}$ are respectively the zero-centered version of $I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}$, the cardinal of $Kset_{x^*}$ and the sample bias of $\hat{f}(x^*)$. Thus $I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}$ takes into account two kinds of uncertainties: the regression's method uncertainty and the observation error. Equation (7.5) shows that the prediction error $\varepsilon^{pred}_{x^*}$ has a normal distribution with the unknown mean $-Bias_{\hat{f}(x^*)}$. The prediction error tolerance interval $I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}$ is constructed based on the $Kset_{x^*}$, which is a finite sample size, so it is centered on the sample bias $-\widehat{bias}_{\hat{f}(x^*)}$. However, because of its definition, $I(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta}$ is guaranteed with confidence level γ, to contain at least a proportion β of the normal distribution of the prediction error at x^*. Hence we have:

$$P_{\mathcal{T}}\left(P_{\varepsilon}\left(L(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta} \leq \varepsilon^{pred}_{x^*} \leq U(\varepsilon^{pred}_{x^*})^T_{\gamma,\beta} \middle| \mathcal{T}\right) \geq \beta\right) = \gamma,$$

where $\mathcal{T} = (\hat{f}(x^*), \sigma_{x^*})$ is the estimated vector at point x^*. This equation can be rewritten

$$\begin{aligned} & P_{\mathcal{T}}\left(P_{\varepsilon}\left(L(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} \leq \varepsilon + f(x^*) - \hat{f}(x^*) \leq U(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} \middle| \mathcal{T}\right) \geq \beta\right) \\ = & P_{\mathcal{T}}\left(P_{\varepsilon}\left(\hat{f}(x^*) + L(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} \leq Y(x^*) \leq \hat{f}(x^*) + U(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} \middle| \mathcal{T}\right) \geq \beta\right) \\ = & P_{\mathcal{T}}\left(P_{\varepsilon}\left(Y(x^*) \in (\hat{f}(x^*) + I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T}) \middle| \mathcal{T}\right) \geq \beta\right) = \gamma. \end{aligned} \tag{7.6}$$

Equation (7.6) means that, by taking our assumptions, the tolerance interval for the response variable is computed by adding the local linear regression estimate to the tolerance interval on the prediction error:

$$I(x^*)_{\gamma,\beta}^{T} = \hat{f}(x^*) + I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} \blacksquare$$

Even though we have a biased prediction, our tolerance interval for $Y(x^*)$ contains the desired proportion of the conditional distribution of the response variable. This is due to the fact that our tolerance intervals on the response variable $I(x^*)_{\gamma,\beta}^{T}$ are computed based on the tolerance intervals on the prediction error $I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T}$. LHNPE conditions assume that the prediction error has a unknown normal distribution with mean and variance being respectively the negative bias and the variance of the prediction error. So for high values of γ and for $\beta > 0.5$, $I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T}$ will contain the true bias. Therefore, adding $I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T}$ to the biased regression estimate will remove the bias and give tolerance intervals that works properly with biased regression estimators.

7.1.2 Computational aspect

By taking advantage of the LHNPE conditions for the local linear estimator, the tolerance interval on the prediction error at the point x^*, described by (7.4), is approximated by the tolerance interval on prediction errors inside its LHNPE neighborhood. The prediction error inside the LHNPE neighborhood of the query point is represented by $Eset_{x^*}$ and it is defined formally as:

$$Eset_{x^*} = \{\varepsilon_i^{pred} | (x_i, Y_i) \in Kset_{x^*}\}, \text{ where } \varepsilon_i^{pred} = Y_i - \hat{f}^{-i}(x_i), \tag{7.7}$$

where $\hat{f}^{-i}(x_i)$ is the local linear estimation without using the i^{th} observation, obtained by (4.24). Note that $Y_i - \hat{f}(x_i)$ is a residual and it depends on the random variable Y_i; however, $Y_i - \hat{f}^{-i}(x_i)$ and Y_i are independent.

Hence, given an input vector x^*, K the number of neighbors in $Eset_{x^*}$, β the desired content and γ the confidence level, the tolerance interval for the prediction error variable

$\varepsilon_{x^*}^{pred}$ is computed by replacing $\hat{\theta}, \hat{\sigma}$ and n in Equations (2.11) and (2.12) which results in:

$$I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} = \hat{\theta} \pm \mathbf{c}\hat{\sigma}, \text{ where } \mathbf{c} = \sqrt{\frac{(K-1)(1+\frac{1}{K})Z^2_{1-\frac{1-\beta}{2}}}{\chi^2_{1-\gamma,K-1}}}, \tag{7.8}$$

$$\hat{\theta} = \bar{\varepsilon}_i^{pred} = K^{-1} \sum_{\varepsilon_i^{-i} \in Eset_{x^*}} \varepsilon_i^{pred} \text{ and } \hat{\sigma}^2 = (K-1)^{-1} \sum_{\varepsilon_i^{pred} \in Eset_{x^*}} (\varepsilon_i^{pred} - \bar{\varepsilon}_i^{pred})^2. \tag{7.9}$$

We propose to take the LHNPE neighborhood as the K-nearest neighbors to the query points where K can be a fixed or a variable number tuned on the dataset. So depending on the LHNPE neighborhood selection method, we have two different methods to obtain tolerance intervals for LLR but both methods require 10-fold cross validation or Leave-One-Out (LOO) errors of the whole training set. We denote this by *error_set*:

$$error\_set = \{\varepsilon_i^{pred} | (x_i, Y_i), i \in (1, \cdots, n)\}, \text{ where } \varepsilon_i^{pred} = Y_i - \hat{f}^{-i}(x_i). \tag{7.10}$$

Algorithm 1 summarizes the required steps for obtaining tolerance intervals for local linear regression.

Algorithm 1 Tolerance Interval for local linear regression

1: **for all** $(x_i, Y_i) \in trainingSet$ **do**
2: $\quad \varepsilon_i^{pred} \leftarrow Y_i - \hat{f}^{-i}(x_i)$
3: $\quad error\_set \leftarrow \{error\_set, \varepsilon_i^{pred}\}$
4: **end for**
5: **for all** $x^* \in testSet$ **do**
6: $\quad fval \leftarrow \hat{f}(x^*)$
7: $\quad Kset_{x^*} \leftarrow$ findToleranceNeighborhood(x^*)
8: $\quad Eset_{x^*} \leftarrow$ error of instances in $Kset_{x^*}$, previously stored in *error_set*
9: $\quad I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T} \leftarrow \beta$-content γ-coverage normal tolerance interval of $Eset_{x^*}$ as in Equations (7.8,7.9).
10: $\quad I(x^*)_{\gamma,\beta}^{T} \leftarrow fval + I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T}$
11: **end for**

7.1.3 LHNPE bandwidth with Fixed K

This method takes the K nearest neighbors of x^* as its LHNPE neighborhood. These neighbors are returned by the function "findToleranceNeighborhood(x^*)". K is a fixed number for all the dataset which is tuned as a hyper-parameter. We denote this interval prediction method for LLR by "fixed K". Once the local linear model has been built and *error_set* has been found on the training set, the computational complexity of interval prediction for a new instance is the same as an evaluation under the local linear regression. More explanation can be found in Section 4.3.2. We select this neighborhood in such a way that it remains inside the regression neighborhood. This condition is respected appropriately all points of the feature space of a dataset. Thus we have to take a LHNPE bandwidth

that coherent on the majority of points in the feature space. In Chapter 9, this condition is always satisfied except in the "Auto" dataset where the LHNPE bandwidth is a bit greater than the Regression bandwidth.

7.1.4 LHNPE bandwidth with variable K

The idea behind this LHNPE bandwidth selection method is to find the "best" LHNPE bandwidth (best K) of each input vector x^*. This method is summarized in Algorithm 2. For a fixed value of β, and for each input vector x^*, the computation begins with an initial value of K, then the β-content γ-coverage normal tolerance interval of errors in $Eset_{x^*}$ defined in (7.7) is calculated. This process is repeated for the same input vector x^* but different values of $K, MIN_K \leq K \leq MAX_K$. Finally, the $I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^{T}$ having the smallest size among the tolerance intervals computed by different values of K (different $Eset_{x^*}$) is chosen as the desired interval and is added to $\hat{f}(x^*)$. This iterative procedure leads us to choose the interval that has the best trade-off between the precision and the uncertainty to contain the response value. The more K increases, the less the local homoscedasticity assumptions match reality and this yields a prediction error variance different from the true one. If we find a variance higher than the true one, it could be partially compensated by the fact that the tolerance interval size decreases when the sample size increases. However, an increase in K may lead us to obtain smaller prediction variance; this issue is controlled by MAX_K. On the contrary, when K is small, the LHNPE conditions are respected but the tolerance interval sizes increase just because the sample size is too small. Thus choosing the value of K that minimizes a fixed β-content γ-coverage tolerance interval ensures that we will have the best trade-off between the faithfulness of the local assumptions (LHNPE conditions) and the required sample size to guarantee the desired β proportion of the response value. The optimal value of K may vary much more on heterogeneous datasets. MIN_K and MAX_K are global limits for the search process. MAX_K stops the search process if the best value for K is not found before. This can occur when increasing the neighborhood, it gets contaminated with instances having smaller predictive errors than the prediction of the query point. In practice, these smaller prediction errors usually belong to a different subspace of the feature space with different error variances and/or prediction error distributions. Therefore these two bounds serve to restrict the search process in a region where it is most likely to contain the best neighborhood of x^*. MAX_K is usually included in the regression neighborhood. However one can take it greater than the regression bandwidth and let our search algorithm (Algorithm 2) find the neighborhood which gives the smallest tolerance interval.

Once the local Linear model has been built and *error_set* has been found on the training set, the computational complexity of interval prediction for a new instance is $(MAX_K - MIN_K)$ times higher than the complexity of an evaluation under the local linear regression. Because from the beginning to the $Kset_{x^*}$-finding step, everything is similar to LLR, then in the interval calculation phase, LLR computes just one value and "Var K." computes $(MAX_K - MIN_K)$ intervals. More explanation on the LLR complexity can be

Algorithm 2 LHNPE neighborhood with variable K

function FINDTOLERANCENEIGHBORHOOD(x^*)
  $IntervalSize_{min} \leftarrow \infty$
  $Kset_{return} \leftarrow \emptyset$
  for all $i \in MIN_K, \ldots, MAX_K$ **do**
    $Kset_{x^*} \leftarrow i$ nearest number of instances $(x_i, Y_i) \in trainingSet$ to x^*
    $Eset_{x^*} \leftarrow \varepsilon_i^{-i}$ of instances in $Kset_{x^*}$ previously computed in *error_set*
    $I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^T \leftarrow \beta$-content γ-coverage normal tolerance interval of $Eset_{x^*}$ as in Equations (7.8,7.9).
    if $size(I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^T) \leq IntervalSize_{min}$ **then**
      $Kset_{return} \leftarrow Kset_{x^*}$
      $IntervalSize_{min} \leftarrow size(I(\varepsilon_{x^*}^{pred})_{\gamma,\beta}^T)$
    end if
  end for
  return $Kset_{return}$
end function

found in Section 4.3.2

7.2 Local Linear Predictive Intervals

This section describes how to use tolerance intervals for local linear regression to obtain predictive interval models. First we describe how the confidence level γ in these tolerance intervals can be used to obtain predictive interval models. Then we see how to find the "best" value of γ that provides predictive interval model with the smallest mean interval size.

7.2.1 Local Linear Predictive Intervals

The β-content predictive interval on the prediction error, denoted by $I(\varepsilon_{x^*}^{pred})_{\beta}^P$, is obtained by finding the predictive intervals hyper-parameters which satisfies the MIP constraint in (6.11). Finally, the β-content predictive interval on the response variable is computed by adding local linear regression estimation to the error predictive interval:

$$I(x^*)_{\beta}^P = \hat{f}(x^*) + I(\varepsilon_{x^*}^{pred})_{\beta}^P.$$

As explained in 6.5.2, regression predictive intervals models have two types of hyper-parameters. This first is the regression method's hyper-parameter. In LLR, it is the bandwidth used for regression and it serves to find the *error_set*. The second type of hyper-parameters are the predictive interval hyper-parameters. These hyper-parameters are (K, γ) or (MIN_K, MAX_K, γ), respectively, for predictive intervals with fixed K and predictive intervals with variable K.

7.2.2 Hyper-parameter Tuning

At this stage, we suppose that the local linear regression bandwidth has been found. The hyper-parameter tuning methods are the same for the fixed K method or the variable K. The only difference is that in variable K, we are looking for the pair (MIN_K, MAX_K) instead of K in fixed K. The Hyper-parameter Tuning reduces to the constraint optimization problem listed below where **all the constraints are hard constraints**. The tuning procedure explained here is similar to the one discussed in 6.5.2 except that in this case, we must also tune the LHNPE neighborhood hyper-parameters.

Optimization problem for fixed K:

$$\begin{aligned}(\gamma, K) = \text{Argmin}(MIS), \text{ where } MIS = \frac{1}{n}\sum_{i=1}^{n} I(\varepsilon_{x_i})_{\gamma,\beta}^{T} \\ \text{Subject to: } 0 < \gamma < 1 \\ 0 < K \leq n \\ MIP_{\mathcal{S},\gamma,\beta} \geq \beta \text{ or } MIP_{\mathcal{S},\gamma,\beta} \geq F_{\beta,n}^{\alpha}\end{aligned}$$

Optimization problem for variable K:

$$\begin{aligned}(\gamma, MIN_K, MAX_K) = \text{Argmin}(MIS), \text{ where } MIS = \frac{1}{n}\sum_{i=1}^{n} I(\varepsilon_{x_i})_{\gamma,\beta}^{T} \\ \text{Subject to: } 0 < \gamma < 1 \\ 0 < MIN_K \leq MAX_K \leq n \\ MIP_{\mathcal{S},\gamma,\beta} \geq \beta \text{ or } MIP_{\mathcal{S},\gamma,\beta} \geq F_{\beta,n}^{\alpha}\end{aligned}$$

Algorithm 3 describes how to tune the predictive interval hyper-parameters for variable K. The algorithm used for the fixed K is almost the same so we do not mention it. In a first attempt, γ is considered as a fixed high value like $\gamma = 0.9$ or $\gamma = 0.99$ and we focus on finding the LHNPE neighborhood hyper-parameter: the hyper-parameter K or the pair (MIN_K, MAX_K). In section 6.5.1, we saw that the variable $MIP_{\mathcal{S},\gamma,\beta}$ defined by (6.6) must on average be equal to $n\beta$. Thus we can select the LHNPE neighborhood hyper-parameter(s) which find(s) intervals that, based on a Leave-One-Out (LOO) or 10-fold cross validation scheme on the training set, satisfies the tuning constraint defined in (6.11) and also have the smallest Mean Interval Size (MIS). Once we have found K or (MIN_K, MAX_K) we search for the smallest value of γ that satisfies the following constraint.

$$\textbf{Tuning Constraint: } MIP_{\mathcal{S},\gamma,\beta} = \beta.$$

A more conservative approach could be to find LHNPE neighborhood hyper-parameter(s) which give(s) the tolerance intervals having the smallest mean interval size and also satisfying:

$$\textbf{Conservative Tuning Constraint: } MIP_{\mathcal{S},\gamma,\beta} \geq t^{-1/2}\beta(1-\beta)Z_{1-\alpha}.$$

where $Z_{1-\alpha}$ is the $(1-\alpha)$-quantile,i.e., 0.95 quantile, of the standard normal distribution. Small neighborhoods result in big tolerance intervals, thus higher coverage. As long as K's value is increased, the mean interval size decreases too. However after a threshold, the mean interval size may increase or change a little bit but the coverage begins to decrease. In practice, we usually evaluate the effectiveness of both methods on datasets, and incorporate our a priori knowledge on the hyper-parameter tuning phase. We may first find K for "Fixed K"(tune the first method) and when it comes to the finding (MIN_K, MAX_K), we can try to choose the $[MIN_K, MAX_K]$ interval in a way to contain the fixed K values found before.

$$MIN_K \leq \text{ fixed } K \leq MAX_K.$$

Once K is found, we try to decrease value of γ, which decreases the mean interval size. Our goal is to have the smallest mean tolerance interval size that satisfies our inclusion constraint. The idea is to fix the neighborhood parameters with the values found in the preceding process and decrease γ. This procedure is repeated as long as the inclusion constraint is satisfied. High values of γ will guarantee constraint (6.11) but the computed intervals can be very large. Note that, with this approach, the value of γ can be less than β and this may happen when the local density of the response variable is quite dense. Based on the new value of γ, we can go to the first step and recalculate new values for the neighborhood hyper-parameter (K or the pair (MIN_K, MAX_K)) and this can be repeated for one or two iterations.

7.2.3 Application with Linear Loess

We saw above, how to compute predictive interval models in the general form of local linear regression. This paragraph briefly reviews an application with the linear loess regression method. Loess is a version of linear polynomial regression that for each query point, takes its K nearest instances in the feature space as its neighborhood. We denote loess's regression bandwidth with K_{loess}. Bandwidth selection and weight calculation in loess are similar to KNN, as explained in 4.3.3. Loess uses a first or second degree polynomial, so Linear loess refers to a loess with a first degree polynomial.

Predictive intervals with Linear loess have three or four hyper-parameters: the linear loess bandwidth K_{loess} and the predictive hyper-parameters which are the confidence level γ and the LHNPE bandwidth. As seen above, (K) and (MIN_K, MAX_K) are respectively the LHNE bandwidth for predictive intervals with fixed K and predictive intervals with variable K. Based on (7.3), we usually have :

$$MAX_K \leq K_{loess} \textbf{ or } K \leq K_{loess}.$$

Algorithm 3 Hyper-parameter tuning for predictive interval with variable K.

function TuneHyper-Params($error\_set, \beta$)
  $\gamma \leftarrow 0.99$ ▷ or $\gamma \leftarrow 0.9$ depending on the dataset
  $(MIN_K, MAX_K) \leftarrow (MIN_{K_0}, MAX_{K_0})$ initial values
  for $iteration = 1..3$ **do**
    $(MIP_{\gamma,\beta}, MIS) \leftarrow$ ComputeOnTrainigSet($\beta, \gamma, MIN_K, MAX_K$)
    $MIS_{min} \leftarrow MIS$.
    while SatisfyConstraint($\beta, MIP_{\mathcal{S},\gamma,\beta}$) and $MIS \leq MIS_{min}$ **do**
      $(MIN_K, MAX_K) \leftarrow (MIN_K \pm somestep_1, MAX_K \pm somestep_2)$
      $MIS_{min} \leftarrow MIS$.
      $(MIP_{\gamma,\beta}, MIS) \leftarrow$ComputeOnTrainigSet($\beta, \gamma, MIN_K, MAX_K$)
    end while
    while SatisfyConstraint($\beta, MIP_{\mathcal{S},\gamma,\beta}$) and $MIS \leq MIS_{min}$ **do**
      $\gamma \leftarrow \gamma - step$
      $MIS_{min} \leftarrow MIS$.
      $(MIP_{\gamma,\beta}, MIS) \leftarrow$ComputeOnTrainigSet($\beta, \gamma, MIN_K, MAX_K$)
    end while
  end for
  return (MIN_K, MAX_K, γ)
end function

function SatisfyConstraint(β, val)
  return $MIP_{\mathcal{S},\gamma,\beta} == \beta$ ▷ or $MIP_{\mathcal{S},\gamma,\beta} \geq t^{-1/2}\beta(1-\beta)Z_{1-\alpha}$
end function

function ComputeOnTrainigSet($\beta, \gamma, MIN_K, MAX_K$)
  $MIP_{\mathcal{S},\gamma,\beta} \leftarrow$ compute this value on the training set by algorithm 2 and using a LOO or 10-fold CV schema .
  $MIS \leftarrow$ compute the mean of interval sizes found in the previous step.
  return $(MIP_{\mathcal{S},\gamma,\beta}, MIS)$
end function

7.3 Relationship with Possibility Distributions

In chapter 3, we have seen how to build different possibility distributions encoding a family of probability distributions that may have generated our sample set. They encode confidence bands, tolerance and prediction intervals. These possibility distributions can also be used in a regression context. Thus the "conditional γ-confidence possibility distribution" will be the maximal specific possibility distribution that bounds, with confidence level γ, simultaneously all inter-quantiles of the unknown conditional distribution $Y(x)$ of the sample set. The "conditional γ-CTP distribution" will be the maximal specific possibility distribution that bounds, with confidence level γ, independently each inter-quantile of $Y(x)$. The "conditional prediction distribution" will be the maximal specific possibility distribution that bounds, on average, independently each inter-quantile of $Y(x)$. These conditional possibility distributions could be applied for possibilitic regression with crisp input and output data. Thus the idea is to find the maximum specific possibility distribution where its alpha-cuts are upper bounds on inter-quantiles of the unknown conditional probability distribution $Y(x)$. Based on Proposition (1) in chapter 3, if we have a method to calculate all β-content inter-quantiles of a conditional unimodal symmetric probability distribution, we can build its conditional possibility distribution.

This chapter, along with the previous one have dealt with different methods of obtaining an upper bound on inter-quantiles of the unknown conditional distribution of the response value. All these methods suppose that the conditional distribution and the error distribution are Gaussian, so they satisfy the unimodal symmetric assumption. Thus we can use the proposed β-tolerance intervals in this chapter for "local linear γ-CTP regression". It requires us to obtain β-content tolerance intervals for local linear regression. However, these intervals must be obtained for all the values of β where $\beta \in [0, 1]$. We have done similar work in [Ghasemi Hamed 12a], called "Possibilistic KNN regression using tolerance intervals". This method proposes a possibility distribution which encodes "simultaneous β-content predictive intervals" (explained in the next chapter). Apart from the three possibilistic regression methods defined in the above paragraph, one can take advantage of any two-sided interval prediction method explained in this thesis to define different types of possibilistic regression.

7.4 Illustrations

This section illustrates the performance of predictive intervals for local linear regression. For the LLR algorithm we used loess with a polynomial of first degree (linear loess). Figure 7.1 illustrates the performances of the following non-linear interval prediction methods on the motorcycle dataset:

- "LS-SVM Conv.": the conventional interval prediction method explained in 5.1.1 obtained with a least-square SVM regression.
- "Loess Conv.": the conventional interval prediction method explained in 5.1.1 obtained

with a Linear Loess regression.

- "Fixed K": local linear predictive intervals with fixed K explained in 7.1.3.
- "Var. K": local linear predictive intervals with variable K explained in 7.1.4.
- "NPQR ": two-sided interval prediction by using two non-parametric quantile regression models [Takeuchi 06] as explained in "Estimates of point-wise interval" of 5.3.3.

"NPQR" consists of a lower $(\frac{1-\beta}{2})$-quantile regression model and an upper $(1-\frac{1-\beta}{2})$-quantile regression model. For instance to obtain 90-predictive intervals with "NPQR", we construct a pair of 0.05-"NPQR" and 0.95-"NPQR" regression models. In this figure we can see how well the methods can fit a complex dataset. The models displayed in Figure 7.1 have the same training set and validation set.

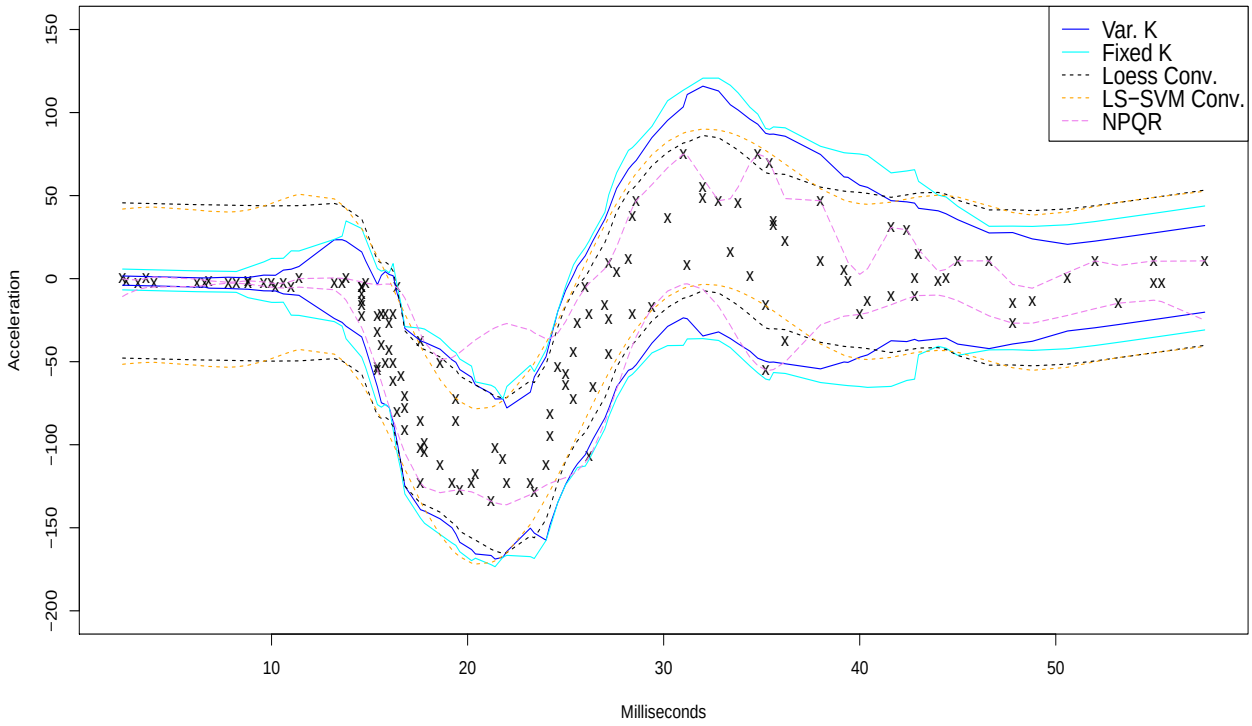

Figure 7.1: Non-linear two-sided 0.95-content interval prediction on motorcycle dataset.

Figure 7.2 shows the same results but with a 10-fold Cross Validation (CV) schema on the whole dataset. The red point displayed by + in Figure 7.2 corresponds to points which the "NPQR " predictive intervals fail to cover. We can see that "NPQR " becomes very unreliable in a CV schema. Thus we added another version of non-parametric quantile regression denoted by "NPQR CV". The "NPQR CV" hyper-parameters are tuned in a way to find intervals that have the smallest MIS and satisfy the MIP tuning constraint in a 10-fold CV on the training set.

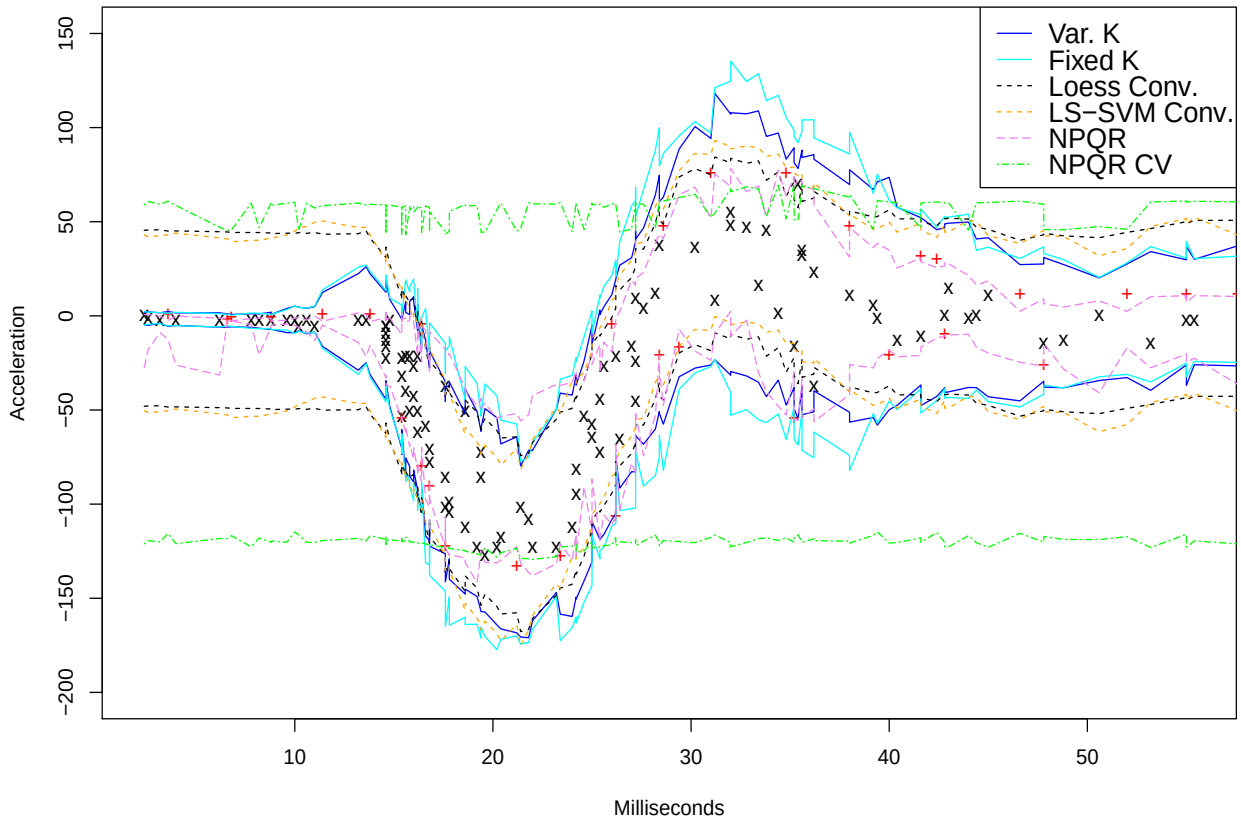

Figure 7.2: Non-linear two-sided 0.95-content interval prediction on motorcycle dataset in a 10-fold cross validation schema.

"NPQR" can describe very precisely the data in Figure 7.1 but the intervals are learned and predicted on the same dataset. However, Figure 7.2 shows that once the model's hyper-parameters are tuned on the training set then tested with 10-fold CV (the "NPQR"

case), the intervals are still small and follow the main structure of the function but they do not contain the desired amount of response value (76.59%) and thus become unreliable. On the other-hand, if we proceed as "NPQR CV" in Figure 7.2, we obtain an acceptable MIP of 94.66 (satisfies the MIP constraint) but it results in very large intervals. The results of this experience are displayed in Table 7.1. We can see that both methods except "NPQR" have their MIP value greater than $F^{0.05}_{0.95,133} = 91.89$, so they satisfy Equation (6.10) and do provide correctly two-sided 0.95-content predictive intervals for this dataset. We can see that our proposed methods have the smallest EGSD values. It means that if all the methods take intervals of the same size, then "Var. K" and "Fixed K" will contain a greater proportion of response values into their predictive intervals than their competitors. While looking at normalized MIS, we observe that "Var. K" also has the smallest value which shows that it gives the tightest reliable envelope. On the other hand, by looking at EGSD and normalized MIS for "NPQR ", we can see that this method is not efficient neither reliable on this dataset.

Interval Prediction Method	MIP	MIS	σ_{is}	normalized MIS	EGSD
LS-SVM Conv.	94.77	93.44	0	0.52	0.52
Loess Conv	93.23	93.46	0	0.52	0.55
Fixed K	97.74	101.66	50.28	0.57	0.48
Var K	**96.31**	**86.77**	**38.65**	**0.49**	**0.45**
NPQR	76.59	62.49	33.55	-	0.57
NPQR CV	94.66	176.65	7.81	1	1

Table 7.1: Experiment results for Figure 7.2.

7.5 Conclusion

This chapter introduced predictive interval models for non-parametric regression. They are applied for two-sided interval predictions but one can also use them in a one-sided interval prediction context. Predictive intervals and predictive interval models were defined in the previous chapter. These models provide intervals which contain at least a desired proportion of the conditional distribution of the response variable given specified combination of predictors. *They can be obtained with tolerance intervals for regression or confidence intervals for regression quantiles but the application of these methods in the non-linear and particularly the non-parametric case are limited in the literature.* The originality of this work is to extend this concept to local linear models. **Our method does not neglect the regression bias and finds intervals that work properly with biased regression models.** For more details see the review of the interval prediction methods in Section 5.4. Finally we have seen that all these methods can also be used for possibilistic regression with crisp input and output data. Our predictive intervals are based on local linear regression (see

[Carroll 88] for a detailed discussion about inference on heteroscedastic regression models). We assume that the mean regression function is locally linear and the prediction error variable $(Y_i - \hat{f}(x_i))$ has locally the same distribution. The idea behind this method is to exploit the local density of prediction error in the LHNPE neighborhood of the query point x^* to find the most appropriate intervals that contain the desired proportion of response values $Y(x^*)$. For this purpose, we use tolerance intervals on prediction errors and they are obtained with a fixed and variable neighborhood method. We use the leave-one-out or 10-fold cross validation errors of the regression function to obtain the predictive intervals. These errors are obtained based on a local linear estimation which could be done by any bandwidth selection technique. **Once the mentioned errors have been found, we can use them to obtain our non-parametric predictive intervals. For this purpose, we need a second bandwidth, which is the tolerance interval bandwidth.** The LHNPE bandwidth is always included in the regression bandwidth. **One must not confuse our predictive intervals with bandwidth selection methods for local polynomial regression.** Local linear regression needs a bandwidth, but is not just a bandwidth selection method. In the same way, our predictive interval method requires a bandwidth on the local prediction errors, but it is not just a bandwidth selection method.

Our method differs from conventional least-squares approaches for finding confidence intervals on the unknown conditional mean function explained in 5.1.2. Because our method takes into account the sample size and finds confidence intervals on inter-quantile of the local distribution for the response variable $f(x) + \varepsilon$, while the conventional methods just estimate asymptotic global inter-quantiles for the conditional response variable (or the conditional mean estimate). Most practitioners of the Machine Learning community usually estimate such predictive intervals by another method described in 5.1.1. We have seen that this method has a very small area of application and does not take into account the sample size. In Chapter 9, we will see that it is one of the most unreliable predictive interval techniques.

Contrary to quantile regression, our method is based on the local linear least squares model, so one can obtain both the conditional mean function and the predictive intervals. Another main difference is that quantile regression obtains estimates of quantiles which, on average, estimate the true quantile function but our method proposes predictive intervals which contain at least a desired proportion of the conditional response variable. Quantile regression may sometimes be more robust than least-squares estimators but it suffers from several problems. One of these problems is the absence of a conditional quantile function. It can occur where the conditional variance of the error distribution is not a function of predictors. Now consider the case when the conditional quantile function is different to the conditional mean function. We know that the conditional mean estimator converges faster than the conditional quantile estimator [Koenker 05]. Thus estimating intervals by quantile regression may be less efficient than using least-squares methods. Besides, it is important to note that quantile regression also suffers from the crossing quantile problem, which is not present here. Our proposed methods are in the class of least-squares based interval prediction methods, so they take advantage of their fast convergence. However they are more reliable and efficient than the other members of this class (conventional methods).

This is because our methods take into account the sample size and find confidence intervals on inter-quantiles of the local distribution for the response variable whereas the conventional methods just estimate asymptotic global inter-quantiles of the conditional response variable. The next chapter discusses "simultaneous predictive intervals with KNN". Then in Chapter 9, our proposed predictive interval models, as well as other conventional interval prediction methods, linear quantile regression, confidence interval on linear regression quantile and a non-linear quantile regression method are applied on nine different benchmark regression datasets. The results show that our approach performs very well. It is significantly more effective than other methods and remains the most reliable non-linear interval prediction method. In the following cases, our method may have **similar results to its alternatives**:

- For interval prediction models which contain a very high proportion (0.99 or more) of the distribution of $Y(x)$.
- The dataset is almost identically distributed in the feature space.
- The dataset has quite low heteroscedasticity.

Our methods are not suited when:

- The reliability of the predicted intervals is not a concern.
- There exits regression models having significantly better prediction results than non-parametric regression models.
- The prediction errors are not normally distributed.

The advantages and drawbacks of our methods are listed below:

Advantages

- It is a reliable interval prediction method for local linear least squares models.
- It does not ignore the non-parametric regression bias.
- It can be used with models having heteroscedastic errors.
- It directly addresses the problem of having predictive intervals that contain at least the desired proportion of response values. It is not designed to work asymptotically and also works with small datasets.
- It does not suffer from the crossing quantiles effect.
- It provides one model for two-sided interval prediction.
- It is simple, reliable and effective.
- It is based on local linear regression, which is a well-known regression method.

Drawbacks

- It is currently just based on local linear regression.
- It has a greater computational complexity than conventional and quantile regression interval prediction methods.

Chapter 8

Simultaneous Predictive Intervals for KNN

Contents

This chapter introduces the concept of simultaneous predictive intervals. These intervals form an envelope around the regression estimate which contains simultaneously a proportion β of the whole distribution of the response variable Y. A simultaneous predictive interval model provides simultaneous predictive intervals for all the points in the predictor space, $\forall x \in \mathcal{X}$. An interval alone cannot be a simultaneous predictive interval because it cannot assure the simultaneous condition (an interval is not an envelope). Thus we use the term "simultaneous predictive interval" to refer to models which provide such intervals. β-content simultaneous predictive intervals can be obtained with simultaneous tolerance intervals for regression in linear regression. This work introduces simultaneous predictive intervals for KNN Regression. This work is similar to predictive intervals with local linear regression but it has three main differences: first, it is performed in a simultaneous context. Second, it uses a KNN regression method instead of a local linear one, and finally the simultaneous predictive interval for the response value is obtained directly with the observation values instead of prediction errors. This chapter briefly discusses these intervals, but the interested reader can find more details in [Ghasemi Hamed 12a] [Ghasemi Hamed 12c].

8.1 Simultaneous Predictive Intervals

Definition 25 *Let $\mathcal{S} = \{(x_1, Y_1) \cdots, (x_n, Y_n)\}$ denote a random sample where $Y_i = f(x_i) + \varepsilon_i$ and ε_i is white noise. β-content simultaneous predictive intervals for x, are denoted by*

$I(x)_\beta^{SP}$, and they are such that:

$$P_{Y(x)}\left(\left(Y(x) \in I(x)_\beta^{SP} \middle| \mathcal{S}\right), \text{ for all } x \in \mathcal{X}\right) \geq \beta, \text{ where } I(x)_\beta^{SP} = [L(x)_\beta^{SP}, U(x)_\beta^{SP}] \quad (8.1)$$

Since we have observed $\mathcal{S}$, $I(x)_\beta^{SP}$ is no longer random, and the probability measure is just related to cover at least a proportion β of the conditional distribution of the response variable $Y(x)$, simultaneously for every $x \in \mathcal{X}$.

Simultaneous predictive intervals contain, for all $x \in \mathcal{X}$, simultaneously, at least a desired proportion of the conditional distribution of the response variable. *Simultaneous predictive intervals for linear models can be obtained by simultaneous tolerance intervals but these concepts have only so far been treated for linear models.* We have addressed a similar concept for non-parametric regression with crisp input and output data [Ghasemi Hamed 12c]. The reader can also find a related study under the possibility theory [Ghasemi Hamed 12a].

8.2 Testing the Models

In this case we use the test defined in (8.2), to verify if a model, built on a dataset, respects the coverage required by simultaneous predictive intervals. *This is done in a* 10*-fold cross validation and we expect the fraction of prediction values inside the envelope to be greater or equal to β, for each of the* 10 *models in cross validation. For example, for $\beta = 0.95$ in a* 10*-fold cross validation, it is expected that each of the* 10 *built models to have a Mean Inclusion Percentage (MIP) greater than or equal to* 0.95. Thus for each fold must satisfy:

$$\textbf{Simultaneous MIP constraint: } MIP_{\mathcal{S},\beta} \geq \beta, \quad (8.2)$$

where $MIP_{\mathcal{S},\beta}$ defined in (6.6) has a different value for each fold; β is the desired simultaneous proportion of the conditional distribution, and n_v is the number of test instances used to validate each of the 10 models. Every concept seen in Chapter 6 for predictive interval models is the same for simultaneous predictive intervals, expect for the MIP constraint. This case is simpler since each fold MIP must be at least βn.

8.3 KNN simultaneous predictive intervals

As in the predictive interval models case, there are two ways to obtain simultaneous predictive interval models: the first employs simultaneous tolerance intervals with a very high confidence level like $\gamma \approx 1$; and the second finds the smallest confidence level $0 < \gamma_0 < 1$ which includes the true unknown model. The concept of simultaneous predictive intervals refers to both intervals. There is not yet any work in the literature for simultaneous tolerance intervals with the KNN regression method, thus we introduced a direct way of obtaining such intervals. This chapter has been published in [Ghasemi Hamed 12c] as "Simultaneous

Interval Regression for K-Nearest Neighbor". Our method exploits the local density of the neighborhood to find the most appropriate intervals to contain the desired proportion of response values, so the proposed interval construction method may be more effective with heterogeneous data set with heteroscedastic error. This method is very similar to local linear predictive intervals with variable K, explained in 7.1.4, except that it is used to obtain a simultaneous content, the regression method is KNN instead of locally linear and the simultaneous predictive interval for the response value is obtained directly with the observation values instead of prediction errors.

The Method

Equation (8.3) gives a γ-confidence β-content normal tolerance interval for the response values of observations that are in the Simultaneous Predictive Intervals (SPI) neighborhood, $Kset_{x^*}$. SPI neighborhood $Kset_{x^*}$ is the sample set that contains the response values of the K-nearest neighbors of x^*. This neighborhood can provide SPI intervals. We suppose that given an input vector x^*, K, β, and γ such SPI neighborhood exits and if we find it along with a sufficiently high confidence level γ, this equation can provide us with the Simultaneous Predictive Intervals (SPI) for the response variable:

$$I(x^*)_{\beta}^{SP} = \hat{\theta} \pm \mathbf{c}\hat{\sigma}, \ \textit{where } \mathbf{c} = \sqrt{\frac{(K-1)(1+\frac{1}{K})Z^2_{1-\frac{1-\beta}{2}}}{\chi^2_{1-\gamma,K-1}}}, \tag{8.3}$$

$$\hat{\theta} = \hat{f}(x^*) = \frac{\sum_{i=1}^{n} \mathcal{K}_b(d(x^*, x_i))Y(x_i)}{\sum_{i=1}^{n} \mathcal{K}_b(d(x^*, x_i))}, \tag{8.4}$$

$$\hat{\sigma}^2 = (K-1)^{-1} \sum_{Y(x_i) \in Kset_{x^*}} (Y(x_i) - \bar{Y})^2 \ \textit{and} \tag{8.5}$$

$$\bar{Y} = K^{-1} \sum_{Y(x_i) \in Kset_{x^*}} Y(x_i). \tag{8.6}$$

The symbol $\hat{f}(x^*)$ above denotes the KNN regression estimate (Equation (4.29)). The SPI neighborhood $Kset_{x^*}$ can be found in two ways: with a fixed number of neighbors (Fixed K) or with a variable number of neighbors (variable K). The fixed K idea in KNN regression comes from the assumption which supposes that the data is homogeneously distributed in the feature space. In this work the constructed intervals are used to find the "best" predictive neighborhood (best K) for each input vector x^*.

Our method is described by Algorithm 4. For a fixed value of β, and for each input vector x^*, the computation begins by an initial value of K, then the β-content γ-coverage normal tolerance interval of $Kset_{x^*}$ is calculated. This process is repeated for the same input vector x^* but different values of $K, MIN_K \leq K \leq MAX_K$. Finally, for a given x^*, the interval having the smallest size between other tolerance intervals for different $Kset_x$, where $MIN_K \leq K \leq MAX_K$, is chosen as the desired interval.

Algorithm 4 KNN simultaneous predictive intervals

for all $x^* \in testSet$ **do**
  $IntervalSize_{min} \leftarrow Inf$
  for all $i \in MIN_K, \ldots, MAX_K$ **do**
    $Kset_{x^*} \leftarrow$ response value of the K nearest instances to x^*
    $I(x^*)^{SP}_{\beta} \leftarrow$ SPI for $Kset_{x^*}$ by Equation (8.3)
    if $size(I(x^*)^{SP}_{\gamma,\beta}) < IntervalSize_{min}$ **then**
      $foundInterval \leftarrow I(x^*)^{SP}_{\beta}$
      $IntervalSize_{min} \leftarrow size(I(x^*)^{SP}_{\beta})$
    end if
  end for
  $I(x^*)^{SP}_{\beta} \leftarrow foundInterval$
end for

How it works

Our variable bandwidth leads us to choose the interval that has the best trade-off between the precision and the uncertainty to contain the response value. Indeed, when K decreases the neighborhood considered is more reliable but it increases the uncertainty of the estimation. On the contrary, when K increases, the neighborhood becomes less reliable but the size of the tolerance intervals decreases. In fact the intervals take into account the number of instances in the neighborhood, and their size also reflects the neighborhood's density. Thus, choosing K that minimizes a fixed β-content γ-coverage normal tolerance ensures the best trade off between the faithfulness of the neighborhood and the uncertainty of the prediction due to the sample size. For the case of the computational complexity, the computation process of KNN simultaneous interval regression is $(MAX_K - MIN_K)$ times higher than the complexity of KNN regression with fixed K. Because from the beginning to the $Kset_x$ finding step, everything remains the same for both regression methods, then in the interval calculation phase, KNN regression with fixed K computes just one interval and instead our method computes MAX_K ones. For more details on the complexity of KNN see [Silverman 86].

MIN_K and MAX_K are global limits for the search process and play a similar role to the local linear predictive intervals with variable K. They restrict the search process in a region where it is most likely to contain the best neighborhood of x. MIN_K, MAX_K and γ are algorithms hyper-parameter and they can be found by evaluating the effectiveness of the algorithm on the training set. Our goal is to find an envelope that gives a proportion to be greater or equal to β of all the predictions. The process of finding the optimal value of γ is like predictive intervals explained in Section 7.2.2. High values of γ will guarantee that $MIP \geq \beta$ but the computed intervals can become very large. Thus, we search for the smallest value of γ that satisfies the simultaneous MIP constraint. Note that, with this approach, the value of γ can be lower than β and this may happen when the local density of the response values is quite dense. The next chapter contains some experiments that

help us to rate the reliability and the efficiency of this method.

8.4 Conclusion

The goal of this chapter was to obtains intervals that simultaneously contain a proportion of at least β of the whole distribution of the response variable. For this purpose we introduced Simultaneous Predictive Intervals for KNN regression. Then we used a test to verify if a model claiming to provide such intervals contains the coverage required by these intervals or not. This test is also used as a constraint in the hyper-parameter tuning procedure. In the next chapter, we take advantage of the reliability test defined here to demonstrate the superiority of our method with other alternatives for simultaneous predictive intervals.

Part III

Experiments

Chapter 9

Evaluation of Predictive Interval Models: benchmark datasets

Contents

Chapter 5 discussed several methods for finding intervals in regression models which contain a desired proportion of the response variable. Then, Chapter 6 introduced the concept of predictive interval models and Chapter 7 proposed two methods to obtain such models for local linear regression. In this chapter we will use several regression datasets to compare our predictive interval method for local linear regression with other interval prediction methods. The selected methods will be tested upon their capacity to provide two-sided β-content predictive interval models. The models are compared for their reliability, efficiency, precision and the tightness of their obtained envelope as described in Chapter 6. Note that we are interested in comparing the mentioned methods regardless of any variable selection or outlier detection preprocessing. This chapter is organized in five sections: the

first section describes our datasets, the second section describes the interval prediction methods that are used in the third section. The fourth explains our experiences on the simultaneous predictive models which have been published in [Ghasemi Hamed 12c].

9.1 Benchmark datasets

In this work we use nine benchmark datasets to validate our suggested methods. These datasets are listed below, where we can find each dataset name in double quotes and its abbreviation in parentheses. Then we mention their numbers of predictor and number of instances, respectively denoted by p and n. Note that some of these datasets have fewer variables than their source because we systematically removed any instances having null values. The "Parkinsons Telemonitoring" dataset [Frank 10] contains two regression variables named "motor_UPDRS" and "total_UPDRS". We considered it as two distinct datasets named "Parkinson1" and "Parkinson2". Each dataset has one of the "motor_UPDRS" or "total_UPDRS" variables.

- "Parkinsons Telemonitoring" [Frank 10] (Parkinson1). We extracted Parkinson1 from the "Parkinsons Telemonitoring" [Frank 10] dataset. It has the "total_UPDRS" variable and does not contain the "motor_UPDRS" variable. $n = 5875, p = 21$.
- "Parkinsons Telemonitoring" [Frank 10] (Parkinson2). We extracted Parkinson2 from the "Parkinsons Telemonitoring" [Frank 10]. It has the "motor_UPDRS" variable and does not contain the "total_UPDRS" variable. $n = 5875, p = 21$.
- "Wine Quality" [Cortez 98] (Wine) (Red Wine). $n = 4898, p = 12$.
- "Concrete Compressive Strength" (Concrete) [Yeh 98]. $n = 1030, p = 9$.
- "Housing" [Frank 10] (Housing). $n = 506, p = 14$.
- "Auto MPG" (Auto) [Frank 10]. $n = 392, p = 8$.
- "CPU" [Frank 10] (CPU). $n = 209, p = 7$.
- "Concrete Slump Test" [Yeh 07] (Slump). $n = 103, p = 10$.
- "Motorcycle" (Motorcycle) [Silverman 85]. $n = 133, p = 1$.

9.2 Interval Prediction Methods

In this section we describe the interval prediction methods used to build predictive interval models. Our experiments are performed with the R programming language. So, we first describe how each tested method is implemented in R. Each method has some general and dataset specific hyper-parameters. General hyper-parameter values are given next to the method name in the listing below, and the dataset specific hyper-parameters values are given in Table 9.1. Note that linear models do not have any hyper-parameters.

9.2.1 Method's Implementation

All the interval prediction methods listed below are explained in Chapter 5, except for our predictive-interval method for local linear regression, which is introduced in Chapter 7. The selected methods are as follows:

- "Fixed K": two-sided predictive interval for linear loess as explained in 7.1 with the fixed K LHNPE neighborhood.
- "Var. K": two-sided predictive interval for linear loess as explained in 7.1 with the variable K LHNPE neighborhood.
- "LQR": two-sided interval prediction with linear quantile regression [Koenker 05]. We used the *rq* and *rq.predict* function in *R*'s *quanterg* package.
- "LQRC" two-sided Bonferroni 0.95-level confidence β-content interval obtained with two different quantile regression models as explained in "Confidence based point-wise inference" of 5.3.3. We used the *rq* and *rq.predict* functions in *R*'s *quanterg* package. We use *predict* with the following arguments: interval="confidence", type="percentile", se="boot", bsmethod= "wild".
- "NPQR": two-sided interval prediction by two non-parametric quantile regression models [Takeuchi 06] as explained in "Estimates of point-wise interval" of 5.3.3. This method's hyper-parameter minimizes the Pin-ball loss function with a 10-fold CV on the training set. This method is implemented by the *kqr* function in *R*'s *kernlab* package. We use *kqr* with the following arguments: kernel="rbfdot", for a radial basis kernel function. We set kpar= "automatic" as the default value for radial basis functions. C=4, the cost regularization parameter is set between 3.8 and 5, depending on the dataset.
- "NPQR CV" : two-sided interval prediction by two non-parametric quantile regression models [Takeuchi 06]. The "NPQR CV" hyper-parameters are tuned in a way to find intervals that, in a 10-fold CV on the training set, have the smallest MIS and satisfy the tuning MIP constraint. We use the *kqr* function in *R*'s *kernlab* package. We use *kqr* with the following arguments: kernel="rbfdot", for a radial basis kernel function. We set kpar= "automatic" as the default value for radial basis functions. C=0.1, the cost regularization parameter is chosen to lie 0.05 and 0.2, depending on the dataset. Satisfying the tuning MIP constraint on the training set requires us to select small values of cost regularization parameters.
- "LS-SVM Conv.": the conventional interval prediction method explained in 5.1.1 obtained with a least-square SVM regression. We used the *ksvm* function in *R*'s *kernlab* package. We use *ksvm* with the following arguments: kernel="rbfdot", for a radial basis kernel function. kpar= list(sigma= 0.2), the sigma hyper-parameter is set between 0.01 and 0.45, depending on the dataset, except for the motorcycle dataset which has sigma=6. We also set tau = 0.01, reduced = TRUE, tol = 0.0001.

- "Loess Conv." the conventional interval prediction method explained in 5.1.1 obtained with a linear loess regression.

We use the Tricube kernel, as in [Cleveland 88], as the kernel function in all of our experiments.

9.2.2 Dataset Specific Hyper-Parameters

The linear Loess regression uses the K_{loess}-nearest neighbors as the bandwidth. This K_{loess} is found by minimizing the 10-fold cross validation error on the training set. For more details about linear loess see 4.3.4. All the non-linear methods listed above have at least one hyper-parameter that must be tuned on the dataset. These hyper-parameters are mentioned in Table 9.1 except for "Fixed K" and "Var. K", because "Fixed K" and "Var. K" may have different hyper-parameter for different β value. Thus their hyper-parameter values are mentioned with their method results.

Dataset	"NPQR" C	"NPQR CV" C	"LS-SVM Conv." sigma	"Loss Conv." K_{loess}
Parkinson1	5	0.2	0.25	80
Parkinson2	5	0.1	0.2	70
Wine	5	0.1	0.45	150
Concrete	4	0.1	0.3	80
Housing	4.5	1	0.08	60
Auto	3.8	0.2	0.25	30
CPU	4	0.2	0.025	40
Slump	4.5	0.05	0.05	30
Motorcycle	4	0.1	6	15

Table 9.1: Hyper-parameter values for non-linear interval prediction models.

Hyper-parameter tuning strategy

In a first attempt, datasets are divided into two subsamples of size $\frac{2}{3}n$ and $\frac{1}{3}n$, where n represents the dataset size. The part containing $\frac{2}{3}$ of observation is used to tune the predictive interval model's hyper-parameters. Then, all of the instances will serve to validate the results using a 10-cross validation scheme. **Note that we are interested in comparing the mentioned methods regardless of any variable selection or outlier detection preprocessing.**

9.3 Testing Predictive Interval Models

The goal of this section is to compare the above-mentioned interval prediction methods based on their strength while providing β-content predictive interval models. The models

are compared based on **reliability, efficiency, precision and the tightness of their envelope**. Our introduced methods ("Var. K" and "Fixed K.") are used to obtain predictive interval models for Local Linear Regression (LLR). Consequently, we first compare our methods with the conventional interval prediction on the local linear regression ("Loess Conv."). For this purpose, we will use Tables 9.2 and 9.3 which compare "Loess Conv.", "Var. K" and "Fixed K.". These models are built upon the same regression model and their only difference is their interval computation algorithm. Our tables provide detailed experimental results but they take a lot of spaces which make them hard to interpret, and not useful for comparing several methods across different datasets. We will use MIP charts, MIS charts and EGSD charts to compare all of the interval prediction methods. This comparison measures a method's strength, while providing β-predictive interval models with $\beta = 0.8, 0.9, 0.95$ and 0.99. We have chosen five big datasets and compare in a very detailed manner the precision, reliability, efficiency and envelope width of our models with the conventional model which is its most efficient competitor.

9.3.1 Comparing Local linear Methods

Outliers, limited number of observations and contrast between our assumptions and the true regression function cause errors in the prediction process. These errors occur in a similar manner when estimating the response variable distribution and they increase with β. For $\beta = 0.9, 0.95$, and particularly for $\beta = 0.99$, it becomes a critical task to find an effective interval prediction procedure that is able to find an upper bound for inter-quantiles of $Y(x)$. However these inter-quantiles are the most used ones in machine-learning and statistical hypothesis-testing. Hence, we will compare the methods based on their strength, while providing β-predictive interval models with $\beta = 0.8, 0.9, 0.95$ and 0.99.

Tables 9.2 and 9.3 are used to display the direct dataset measures explained in Chapter 6, for each dataset. These tables compare models of "Loess Conv." , "Var K." and "Fixed K.". For each dataset, we have 12 models, (3 methods : "Loess Conv." , "Var K." and "Fixed K." $\times$ 4 β's value : $0.8, 0.9, 0.95$ and $0, 99$). These 12 models are built on the same regression model which is a linear Loess model with K_{loess} as its bandwidth. K_{loess} is represented next to the dataset's name and it is found by minimizing the 10-fold cross validation error on the training set. Then we will use charts to compare our local linear predictive interval models with the other methods.

Table description

In Tables 9.2 and 9.3, each combination of dataset and β has a cell which displays $F^{0.05}_{\beta,n}$ for the underlying experiment. We can see if a model satisfies its MIP test or not. If it does not satisfy this constraint a $\blacktriangleleft$ or $\lhd$ sign may appear. The $\blacktriangleleft$ sign appears when the current model is the only one to fail the MIP test. When more than one of the three compared model fails, the $\lhd$ sign is put near their results. For each experiment, the model which passes the MIP test and has the smallest MIS is distinguished with the * sign. If a method

receives the * sign for two consecutive β of the same dataset, it is annotated in bold and with ***** . When it comes to our introduced model hyper-parameters, "Var K." needs the value of MIN_K, MAX_K and γ and "Fixed K." has a proper value for its K and γ. These hyper-parameters are illustrated in each dataset row.

Table commentaries

By looking at Tables 9.2 and 9.3, one can see that the three methods work for $\beta = 0.8$ on benchmark dataset. When the desired proportion is 0.8, "Var K." is slightly more effective than "Loess Conv." and "Fixed K." finds the biggest intervals. When it comes to $\beta = 0.9$, "Loess Conv." loses its reliability and fails to satisfy the MIP constraint on three datasets. If we increase the desired proportion to 0.95, the situation stays the same for "Var K." and "Fixed K.", but "Loess Conv." becomes much more unreliable. In fact it fails to satisfy the MIP constraint for five of the nine datasets. When looking in more detail, one can observe that "Fixed K." has almost everywhere larger MIP and gives wider intervals than others. It is important to emphasize that **the conventional method is nowhere more reliable than our methods.** We can also observe that "Var K." usually appears with the * sign and it is the only method which becomes bold. It means that it usually works and obtains the tightest band.

9.3.2 Comparing All Methods by Charts

Our tables are not useful for displaying the eight methods listed in 9.2.1, so for the sake of readability we produced Figures 9.1, 9.2, 9.3 and 9.4. These figures are MIP charts for our experiments. We can see that our introduced methods obtain high MIP, but we need more information to compare their reliability and efficiency. For this purpose we will use the MIS ratio charts and EGSD charts that are explained in 6.4.3.

Chart description

Each β value has a MIP, an EGSD and a MIS ratio chart. For each β, its EGSD chart is displayed just after its MIS ratio chart. For example, Figure 9.5 is the MIS Ratio chart for $\beta = 0.8$ and just after Figure 9.6 is the EGSD chart for $\beta = 0.8$. The MIS ratio charts display the MIS ratio for the reliable models (models which pass the MIP test). For a given dataset, the method having the smallest MIS ratio value is that which finds the tightest reliable envelope (the set of all obtained intervals). The EGSD chart displays the normalized EGSD value for all models. For a given dataset, the model having the smallest EGSD value has an Equivalent Gaussian distribution with the smallest variance.

Dataset	Method	80%			90%		
		MIP	$MIS\ (\sigma_{is})$	$F^{0.05}_{0.8,n}$	MIP	$MIS\ (\sigma_{is})$	$F^{0.05}_{0.9,n}$
Parkinson1 (n=5875, p=21), K_{loess}=80	Loess Conv.	86.99	4.96	79.14	90.08	6.37	89.35
	Fixed K = 40, $\gamma = 0.9$	91.55	5.48 (4.4)		94.88	7.04 (5.64)	
	Var. K *	88.55	4.39 (3.78) *		92.81	5.64 (4.85) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20\ ,\ 60\ ,\ 0.9)$					
Parkinson2 (n=5875, p=21), K_{loess}=70	Loess Conv.	86.36	3.53	79.14	89.95	4.53	89.35
	Fixed K = 50, $\gamma = 0.9$	91.46	4.2 (3.22)		94.64	5.4 (4.14)	
	Var. K *	89.08	3.52 (2.95) *		93.03	4.52 (3.79) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 60\ ,\ 0.9)$					
Wine (n=4898, p=12), K_{loess}=150	Loess Conv.	80.45	1.58 *	79.05	88.39 ◀	2.03	89.29
	Fixed K = 50, $\gamma = 0.7$	82.88	1.75 (0.38)		90.62	2.25 (0.48) *	
	Var. K	83.19	1.77 (0.39)		91.09	2.27 (0.51)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 60\ ,\ 0.9)$					
Concrete (n=1030, p=9), K_{loess}=80	Loess Conv.	79.89	16.63 *	77.94	87.56 ◀	21.35	88.46
	Fixed K = 35, $\gamma = 0.5$	82.61	16.76 (5.73)		91.45	21.52 (7.36) *	
	Var. K	83.68	17.03 (5.91)		93	22.2 (7.59)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20\ ,\ 60\ ,\ 0.9)$					
Housing (n=506, p=14), K_{loess}=60	Loess Conv.	87.17	8.47	76.67	92.68	10.88	87.5
	Fixed K = 40, $\gamma = 0.9$	87.97	8.67 (3.31)		92.7	11.14 (4.25)	
	Var. K *	84.59	7.8 (2.8) *		91.72	10.01 (3.6) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 55\ ,\ 0.9)$					
Auto (n=392, p=8), K_{loess}=30	Loess Conv.	89.29	8.56	77.07	93.61	10.98	87.8
	Fixed K = 50, $\gamma = 0.9$	88.27	7.76 (3.14)		94.41	9.96 (4.03)	
	Var. K *	84.7	6.91 (2.81) *		92.61	8.87 (3.61) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 60\ ,\ 0.9)$					
CPU (n=209, p=7), K_{loess}=40	Loess Conv.	79.87	84.64	75.44	85.13 ◀	108.63	86.58
	Fixed K = 40, $\gamma = 0.9$	85.16	88.07 (64.23)		91.4	113.04 (82.44)	
	Var. K *	80.37	78.49 (59.2) *		88.97	100.75 (20.89) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20\ ,\ 50\ ,\ 0.9)$					
Slump (n=103, p=10), K_{loess}=30	Loess Conv.	87.63	5.58	73.51	91.45	7.17	85.13
	Fixed K = 20, $\gamma = 0.5$	85.72	4.85 (1.41)		88.54	6.23 (1.81)	
	Var. K *	83.81	4.32 (1.24) *		87.63	5.55 (1.6) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (15\ ,\ 30\ ,\ 0.5)$					
Motorcycle (n=133, p=1), K_{loess}=30	Loess Conv.	82.57	61.11	74.29	90.21	78.43	85.72
	Fixed K = 35, $\gamma = 0.7$	85.72	66.47 (32.87)		96.31	85.31 (42.2)	
	Var. K *	85.6	56.73 (25.27) *		94	72.82 (32.44) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (15\ ,\ 35\ ,\ 0.7)$					

Table 9.2: Predictive interval models for local linear regression built on benchmark datasets with $\beta = 0.9, \beta = 0.9$.

Dataset	Method	95%			99%		
		MIP	$MIS\ (\sigma_{is})$	$F^{0.05}_{0.95,n}$	MIP	$MIS\ (\sigma_{is})$	$F^{0.05}_{0.99,n}$
Parkinson1 (n=5875, p=21), K_{loess}=80	Loess Conv.	92.35 ◀	7.59	94.53	94.85 ◁	9.97	98.78
	Fixed K = 40, $\gamma = 0.99$	97.61	9.63 (7.7)		98.74 ◁	12.66 (10.2)	
	Var. K	96.31	7.72 (6.51) *		98.08 ◁	10.15 (8.56)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20\ ,\ 60\ ,\ 0.99)$					
Parkinson2 (n=5875, p=21), K_{loess}=70	Loess Conv.	91.91 ◀	5.4	94.53	94.41 ◁	7.09	98.78
	Fixed K = 50, $\gamma = 0.99$	97.4	7.26 (5.57)		98.64 ◁	9.54 (7.32)	
	Var. K	96.35	6.1 (5.07) *		98.13 ◁	8.02 (6.76)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 60\ ,\ 0.99)$					
Wine (n=4898, p=12), K_{loess}=150	Loess Conv.	92.4 ◀	2.43	94.48	97.3 ◁	3.19	98.76
	Fixed K = 50, $\gamma = 0.9$	95.83	2.91 (0.63) *		98.75 ◁	3.82 (0.83)	
	Var. K	96.42	3.05 (0.67)		98.93	4.02 (0.88) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 60\ ,\ 0.9)$					
Concrete (n=1030, p=9), K_{loess}=80	Loess Conv.	93.87 ◀	25.44	93.88	98.53	33.43 *	98.49
	Fixed K = 35, $\gamma = 0.5$	95.62	25.64 (8.77) *		99.02	33.7 (11.53)	
	Var. K	95.72	26.46 (9.04)		99.02	34.77 (11.88)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20\ ,\ 60\ ,\ 0.9)$					
Housing (n=506, p=14), K_{loess}=60	Loess Conv.	95.24	12.96 *	93.18	97.62 ◀	17.04	98.17
	Fixed K = 40, $\gamma = 0.9$	95.45	13.27 (5.07)		98.61	17.44 (6.66) *	
	Var. K	96.24	13.8 (5.01)		98.61	18.14 (6.58)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 50\ ,\ 0.99)$					
Auto (n=392, p=8), K_{loess}=30	Loess Conv.	96.17	13.09	93.4	97.46 ◀	17.2	98.27
	Fixed K = 50, $\gamma = 0.99$	97.2	13.39 (5.42)		98.71	17.6 (7.12)	
	Var. K *	96.44	11.99 (4.82) *		98.71	15.76 (6.34) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (30\ ,\ 60\ ,\ 0.99)$					
CPU (n=209, p=7), K_{loess}=40	Loess Conv.	86.11 ◀	129.45	92.52	91.39 ◁	170.12	97.86
	Fixed K = 40, $\gamma = 0.99$	96.16	154.67 (112.8)		98.07	203.27 (148.24) *	
	Var. K	94.25	137.68 (101.75) *		96.64 ◁	180.95 (133.72)	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20\ ,\ 50\ ,\ 0.99)$					
Slump (n=103, p=10), K_{loess}=30	Loess Conv.	94.36	8.54	91.46	97.18 ◀	11.23	97.38
	Fixed K = 20, $\gamma = 0.9$	97.18	9.35 (2.72)		98.09	12.29 (3.57)	
	Var. K *	96.27	8.16 (2.25) *		98.09	10.73 (2.96) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (15\ ,\ 30\ ,\ 0.9)$					
Motorcycle (n=133, p=1), K_{loess}=30	Loess Conv.	93.23	93.46	91.89	98.51	122.82	97.58
	Fixed K = 35, $\gamma = 0.7$	97.8	101.66 (50.28)		99.23	133.6 (66.08)	
	Var. K *	96.31	86.77 (38.65) *		99.23	114.03 (50.8) *	
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (15\ ,\ 35\ ,\ 0.7)$					

Table 9.3: Predictive interval models for local linear regression built on benchmark datasets with $\beta = 0.95, \beta = 0.99$.

Chart commentaries

Now we can easily compare all the interval prediction methods. Figures 9.5 and 9.6 respectively display the MIS ratio chart and EGSD chart for $\beta = 0.8$. $\beta = 0.8$ is the easiest case and all the methods can provide reliable predictive interval model. One can observe that "Var. K" and "Fixed K" models are almost always more efficient than the others. If we look in more detail, we can see that "Var. K" usually finds both the smallest MIS ratio and EGSD value. The conventional methods "Loess Conv." and "LS-SVM Conv." are the next efficient ones. When it comes to testing $\beta = 0.9$, the situation stays almost the same for "Var. K" and "Fixed K" and "Var. K" remains the most efficient method. Conversely both the conventional methods fail to provide reliable predictive interval model for three of the nine datasets.

Figures 9.9 and 9.10 ($\beta = 0.95$) show that the conventional pair ("Loess Conv." and "LS-SVM Conv.") are definitely not reliable. Their non-working models find wider and less efficient envelope than our proposed models "Var. K" and "Fixed K". The scenario is still the same for "Var. K": It is the method which usually finds the tightest reliable envelope. It also provides models that, even compared to non-reliable models, have the smallest variance of prediction error. Finally let us look at Figures 9.11 and 9.12 ($\beta = 0.99$). In this case "Fixed K" and "LQRC" are the most reliable models. "Fixed K" takes second place. It fails once more than "Fixed K" and "LQRC". When comparing the efficiency, "Var. K" is still the most efficient solution but its gap decreases with others. In this case "Fixed K" becomes approximately as efficient as "Var. K". It is also interesting to note that, for $\beta = 0.99$, "LQRC" provides more efficient models than before.

Note also that both the conventional pair and the "NPQR CV" method fails more for large datasets than for small datasets. Small datasets do not have sufficient observations to reject the null hypothesis, which states that the tested model is a predictive interval model, so we accept their models as predictive interval models. It is also interesting to observe that "NPQR" fails in all cases. Its intervals are neither reliable nor efficient. These results are summarized in Table 9.4. Table 9.4 summarizes all the displayed charts. Each row of this table is dedicated to a different dataset which summarizes three qualities through $\beta = 0.8, 0.9, 0.95$ and $0, 99$. The first quality is the reliability: we cite the method which is the most reliable through $\beta = 0.8, 0.9, 0.95$. The second quality (the third column) shows the method that, for each dataset, generally provides the tightest reliable band and the fourth column displays the most efficient method. In the fourth column we ignore the method's reliability and we just compare its EGSD normalized value with others EGSD normalized value.

Dataset	Most Reliable	Tightest reliable band	General Efficiency (ignore the reliability)
Parkinson1	LQRC	Var. K	Var. K
Parkinson2	LQRC	Var. K	Var. K
Wine	Var. K & LQRC	Fixed K	Loess Conv
Concrete	Var K, Fixed K, LQRC, NPQR CV & Loess Conv	Fixed K.	Fixed K
Housing	Fixed K & LQRC	Fixed K	Fixed K
Auto	Var K, Fixed K, LQRC, NPQR CV & LS-SVM Conv	LS-SVM Conv	LS-SVM Conv
CPU	Fixed K	Var K	Var K
Slump	Var K, Fixed K, NPQR CV, Loess Conv & LS-SVM Conv	Var K	Var K
Motorcycle	Fixed K, Var K, NPQR CV, LS-SVM Conv & Loess Conv	Var K	Var K

Table 9.4: General ranking based on the MIP charts, MIS charts and EGSD charts for $\beta = 0.8, 0.9, 0.95$ and 0.99.

9.3.3 Detailed Comparison Using Plots

In the previous experiments, we concluded that our proposed predictive interval methods are the most reliable and effective method. Our goal is to compare in a very detailed manner the precision, reliability, efficiency and envelope tightness of our methods with one of its most efficient competitors. For this purpose, we have chosen the five largest datasets, because bigger datasets can provide more significant results. If we compare our introduced methods with the most reliable method, we have to select "LQRC". However "LQRC" is not more reliable than "Fixed K" but we have seen that "LQRC" is considerably less efficient and it only begins to be useful for $\beta > 0.95$. Therefore, we will have a more detailed comparison of our methods with the most effective interval prediction methods. We have seen that "Loess Conv." and then "LS-SVM Conv." are the most effective solutions after "Var K.". They have the same interval prediction methods but they use different regression algorithms, so we select "Loess Conv." which is revealed to be a bit more reliable and effective than "LS-SVM Conv." on the largest datasets. For this purpose we will use EGSD plots and MIP plots (described in 6.4.3) to compare "Var K.", "Fixed K." and "Loess Conv.".

Plot interpretation

For each dataset, the EGSD plot compares the efficiency of the tested models and the method having the highest line in this plot is the most inefficient one. MIS plots compares the envelope wideness of reliable models. The method having the most bottom line provides the most reliable envelope in the MIS plot. Note that in the MIS plot each model is plotted

until its failure MIP. Once we have compared methods based on their envelope size and their efficiency, the MIP plot will help us to compare the precision and reliability of interval prediction models. The best model in this plot is the one that has the nearest line to the upper side of the "Nominal MIP line". For further explanation of these plots, see 6.4.3.

Plot commentaries

By looking at Figure 9.13, we can see that when the nominal MIP is greater than or equal to 0.6, the "Loess Conv." model loses its efficiency but Figure 9.14 shows that the "Var K." model is the tightest reliable model. At the same time, Figure 9.15 compares their failure MIP. We can see that when the "Loess Conv." failure MIP is 0.93, our methods give a failure MIP equal to 0.99. This figure also shows that "Var K." has the most precise model and "Loess Conv." has the least reliable and precise one. The same experiment is performed for the Parkinson2 dataset which gives Figures 9.16, 9.17 and 9.18. In this case, "Loess Conv." becomes more efficient than "Fixed K." but it is always less efficient than "Var K.". Then Figures 9.17, shows that the "Var K." method provides again the most tightest reliable band. Next we look at the Parkinson2's MIP plot and we can see that the "Var K." model remains the most precise model. Figures 9.18 states that "Loess Conv." has a failure MIP of 0.93 and it is again the most unreliable solution.

In the Concrete dataset, the previous ranking gets better for "Fixed K.". "Loess Conv." is no longer the most efficient solution. Figures 9.19, 9.20 and 9.18 show that the "Fixed K." model is more effective and more precise than the "Loess Conv." model. The "Fixed K." method provides a model for the concrete dataset that obtains the tightest reliable band. Note that for $NominalMIP \geq 80$ ($\beta \geq 0.8$), its envelope is even tighter than the "Var K.". Our experiments continue with the Wine dataset where the "Loess Conv." model is the most efficient and provides the tightest band, however it has a failure MIP of 0.83 compared to a failure MIP of 0.99 for "Fixed K" and 0.97 for "Var K.".

We finalize our experiments with the Housing dataset where EGSD, MIS and MIP plots are displayed respectively in Figures 9.25, 9.26 and 9.27. Figures 9.25 shows that the "Var K." model, the "Loess Conv." model and the "Fixed K." model are respectively ranked as the first, the second and the third ranking efficient models. "Var K." is again the method that provides the tightest reliable band and for $\beta \geq 0.8$ the "Loess Conv." model is tighter than the "Fixed K." model. Figure 9.27 gives the same ranking for their precision. These rankings are summarized in Table 9.5. Each row of this table is dedicated to a different dataset which summarizes four qualities through 16 different inter-quantiles: $0.25 \leq \beta \leq 0.99$. The first three columns are similar to Table 9.4 except that they are obtained for $0.25 \leq \beta \leq 0.99$. The fourth column displays the method which is generally the most precise. This is the method that its MIP line, compared to other methods, remains the nearest to the upper side of the "Nominal MIP line".

Dataset	Most Reliable	Tightest reliable band	General Efficiency (ignore the reliability)	General Precision
Parkinson1	Var. K & Fixed K.	Var. K	Var. K	Var. K
Parkinson2	Var. K & Fixed K.	Var. K	Var. K	Var. K
Wine	Fixed K.	Loess Conv. for $\beta \leq 0.8$	Loess Conv.	Var. K
Concrete	Var. K & Fixed K.	Fixed K.	Var. K	Var. K.
Housing	Var. K & Fixed K.	Var. K	Var. K	Var. K

Table 9.5: General ranking based on the MIP plots, MIS plots and EGSD plots for $0.25 \leq \beta \leq 0.99$.

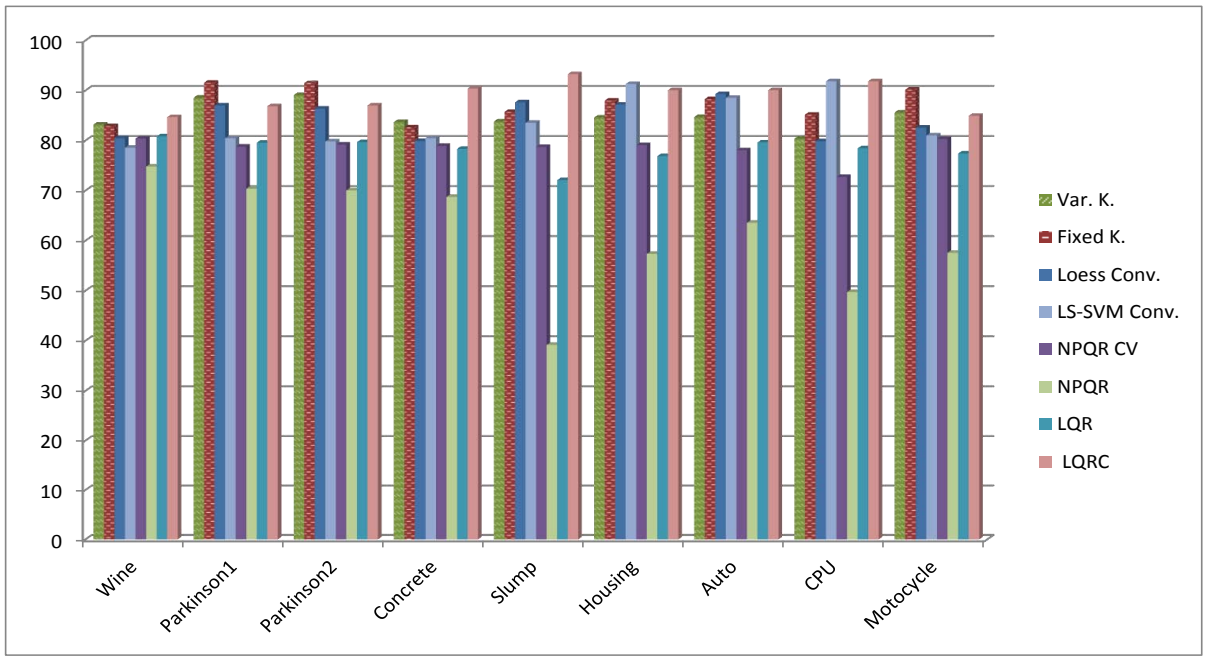

Figure 9.1: MIP chart for benchmark datasets with $\beta = 0.8$.

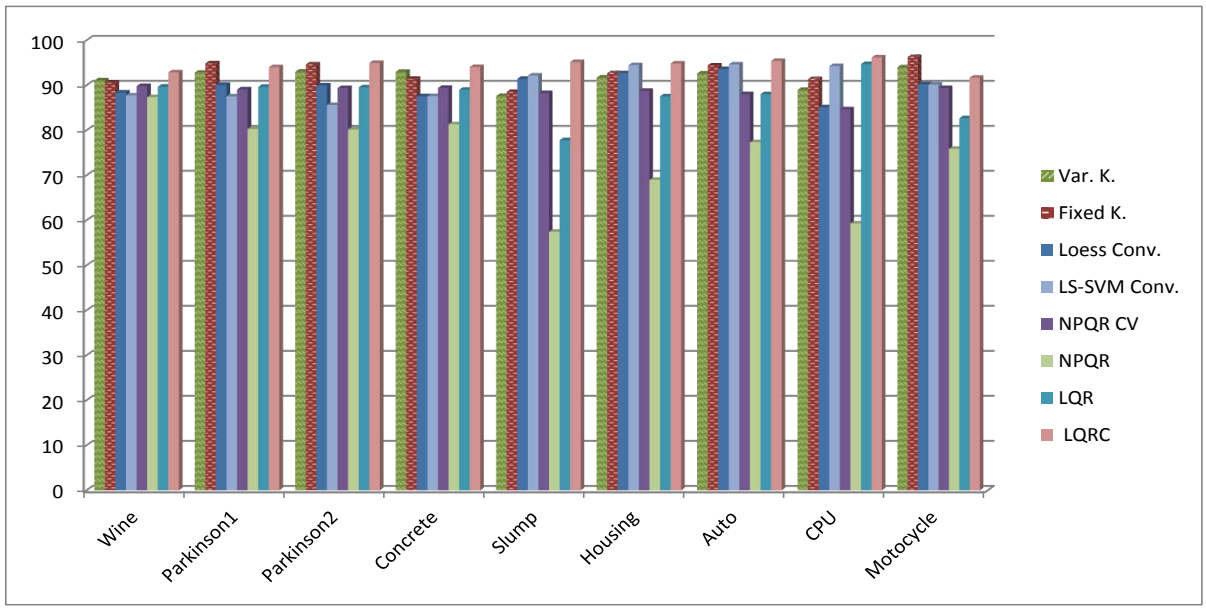

Figure 9.2: MIP chart for benchmark datasets with $\beta = 0.9$.

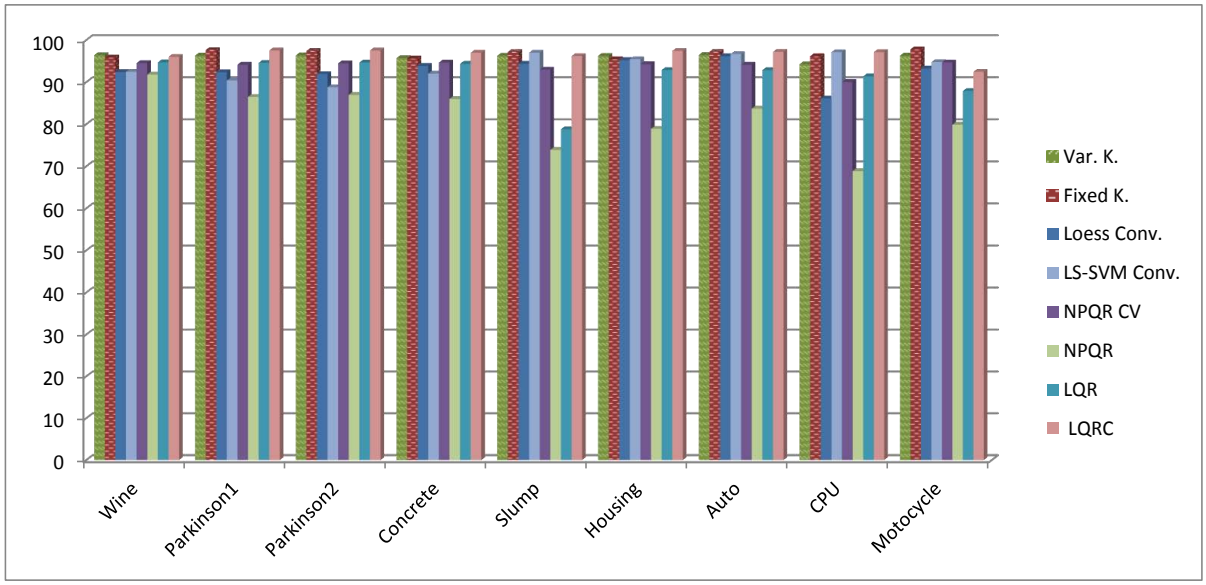

Figure 9.3: MIP chart for benchmark datasets with $\beta = 0.95$.

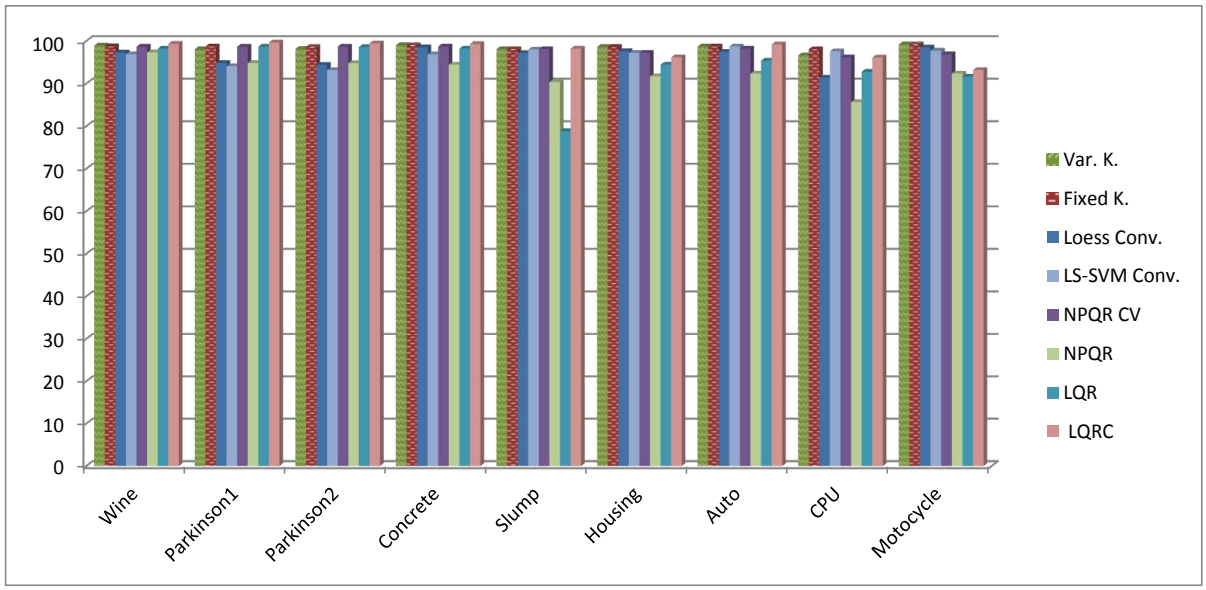

Figure 9.4: MIP chart for benchmark datasets with $\beta = 0.99$.

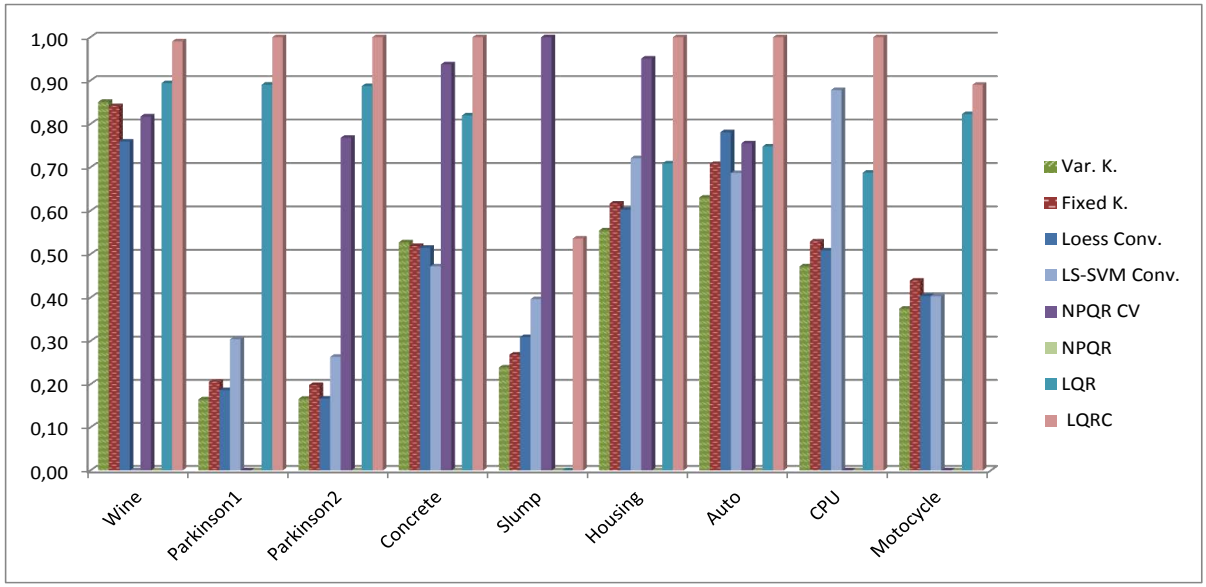

Figure 9.5: MIS Ratio chart for benchmark datasets with $\beta = 0.8$. The smallest value denotes the tightest reliable band.

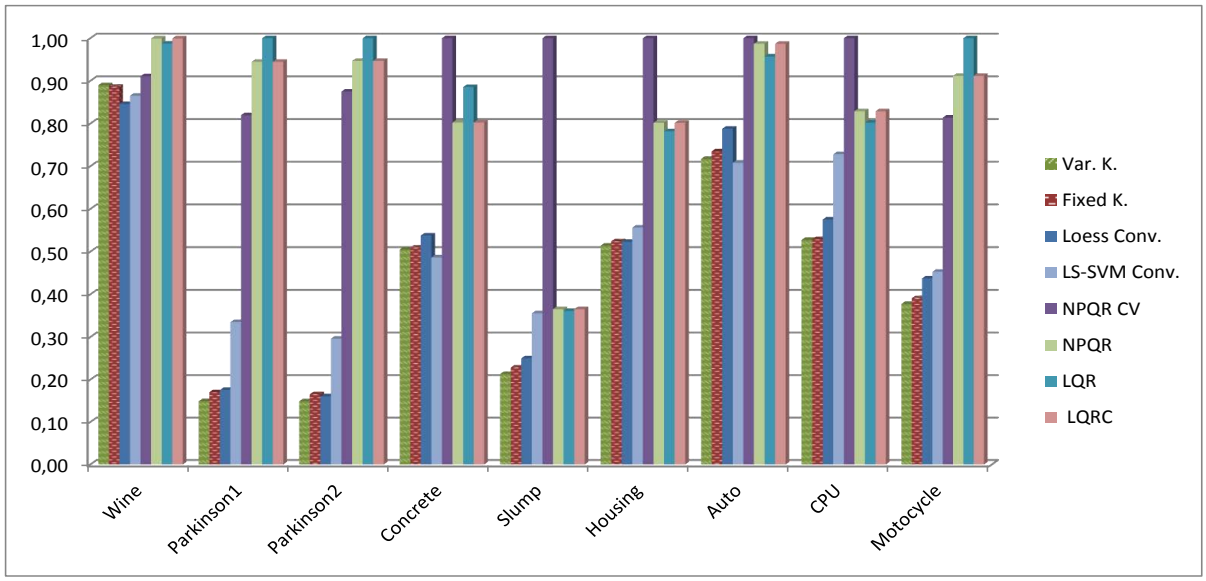

Figure 9.6: EGSD chart for benchmark datasets with $\beta = 0.8$. The smallest value denotes the most efficient band. This measure ignores the reliability.

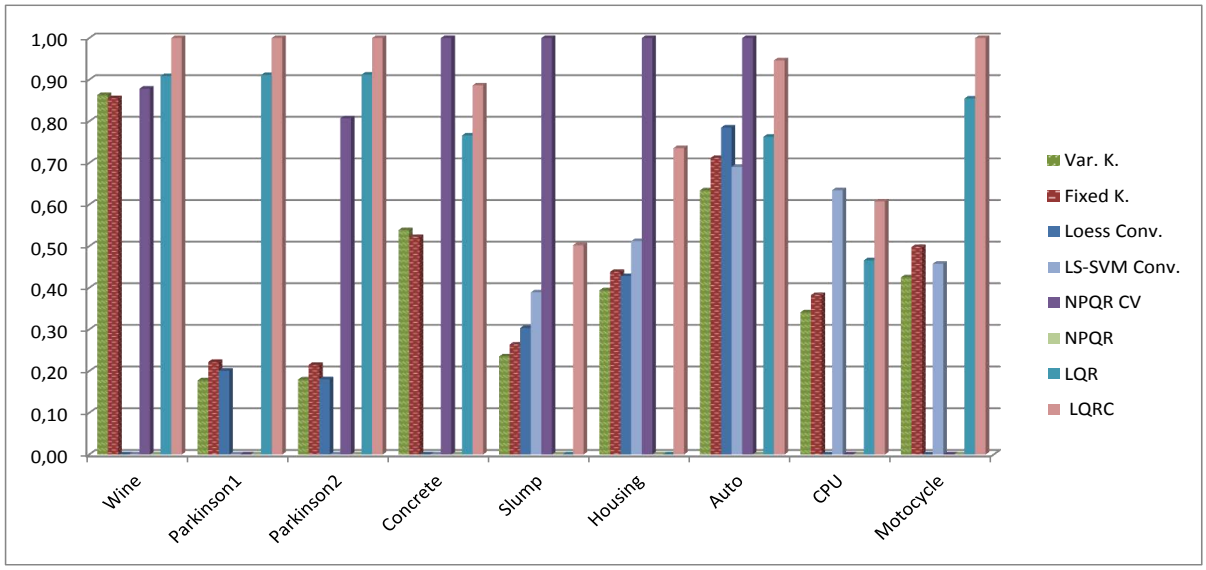

Figure 9.7: MIS Ratio chart for benchmark datasets with $\beta = 0.9$. The smallest value denotes the tightest reliable band.

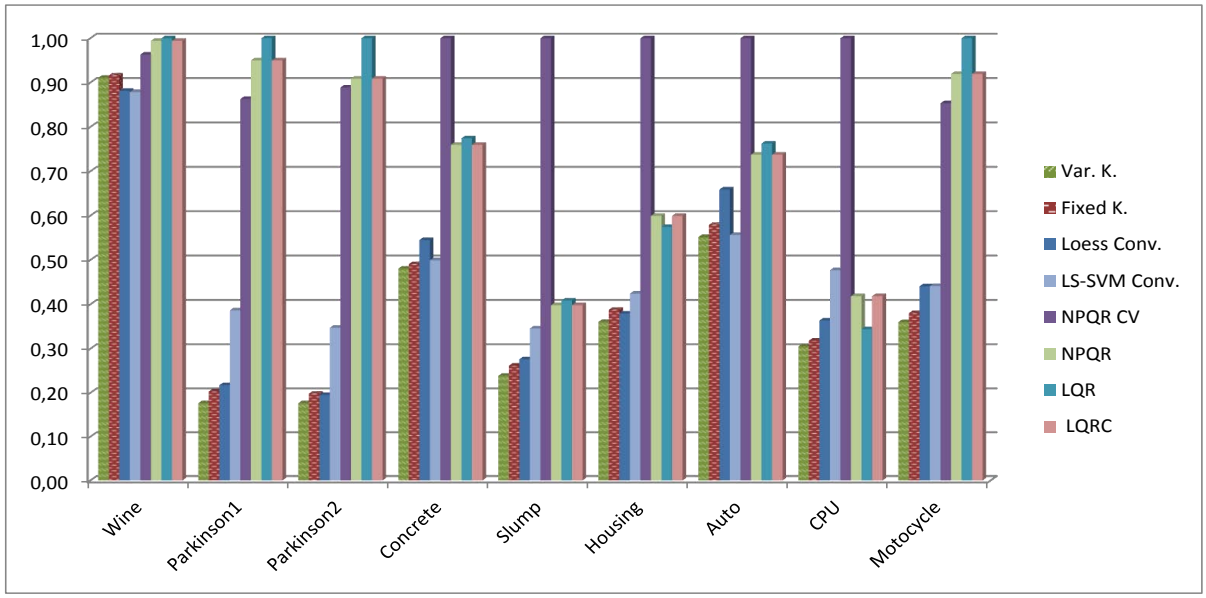

Figure 9.8: EGSD chart for benchmark datasets with $\beta = 0.9$. The smallest value denotes the most efficient band. This measure ignores the reliability.

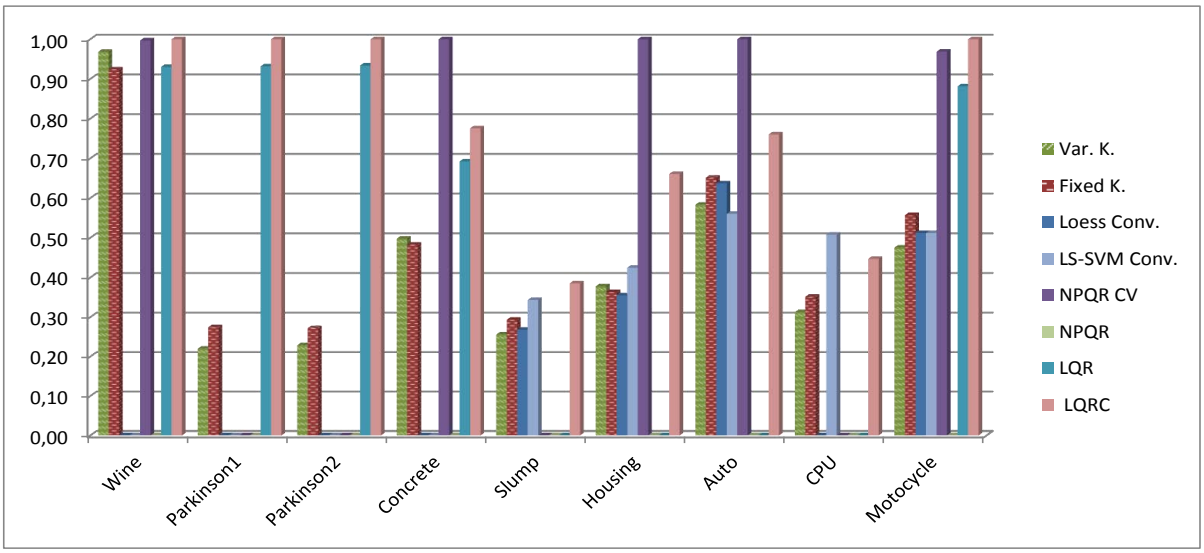

Figure 9.9: MIS Ratio chart for benchmark datasets with $\beta = 0.95$. The smallest value denotes the tightest reliable band.

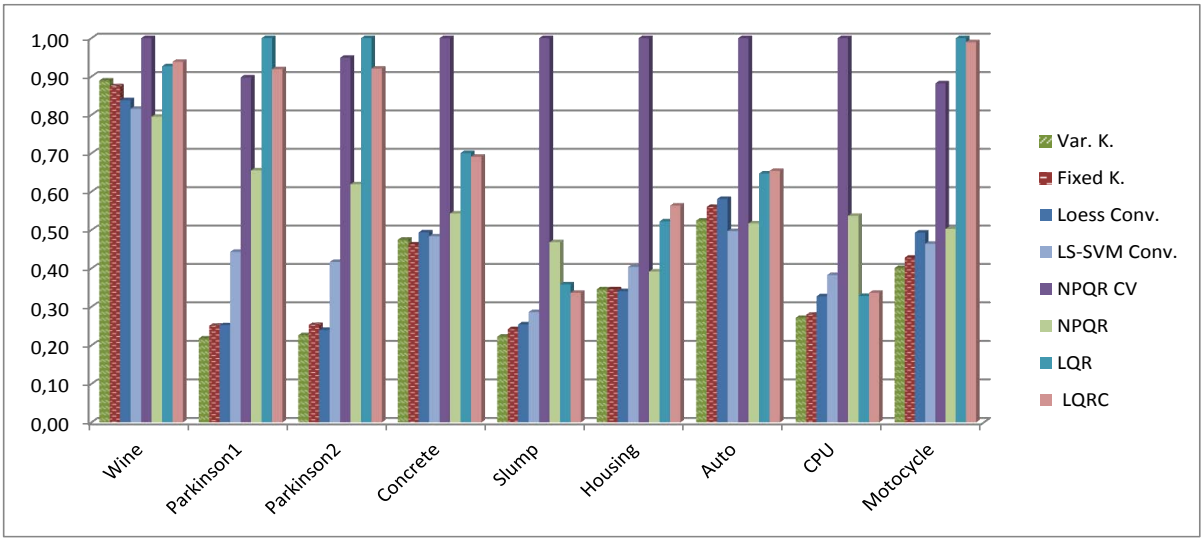

Figure 9.10: EGSD chart for benchmark datasets with $\beta = 0.95$. The smallest value denotes the most efficient band. This measure ignores the reliability.

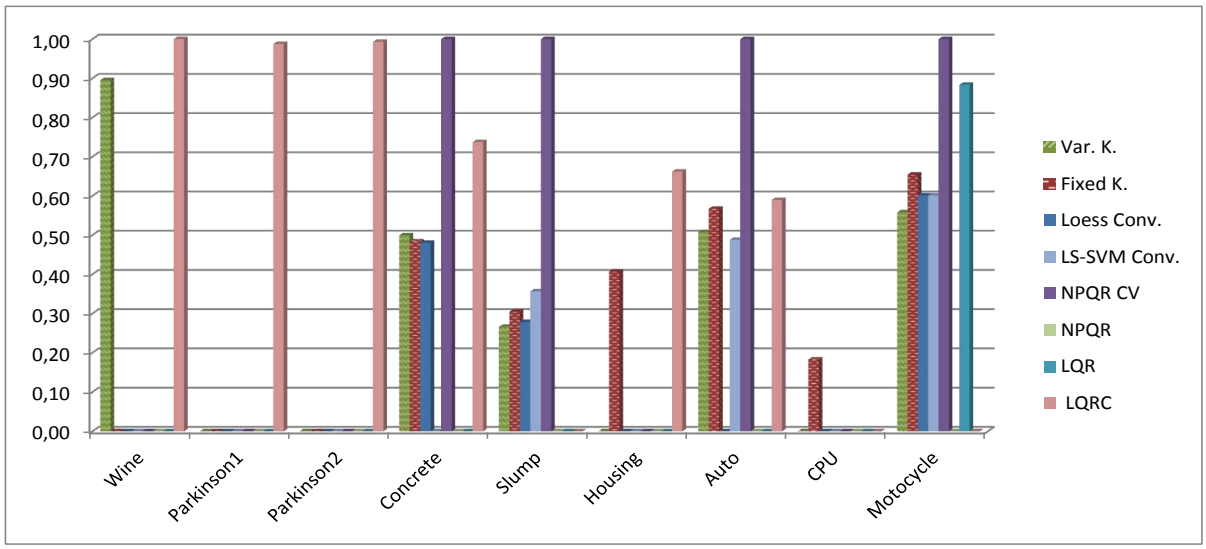

Figure 9.11: MIS Ratio chart for benchmark datasets with $\beta = 0.99$. The smallest value denotes the tightest reliable band.

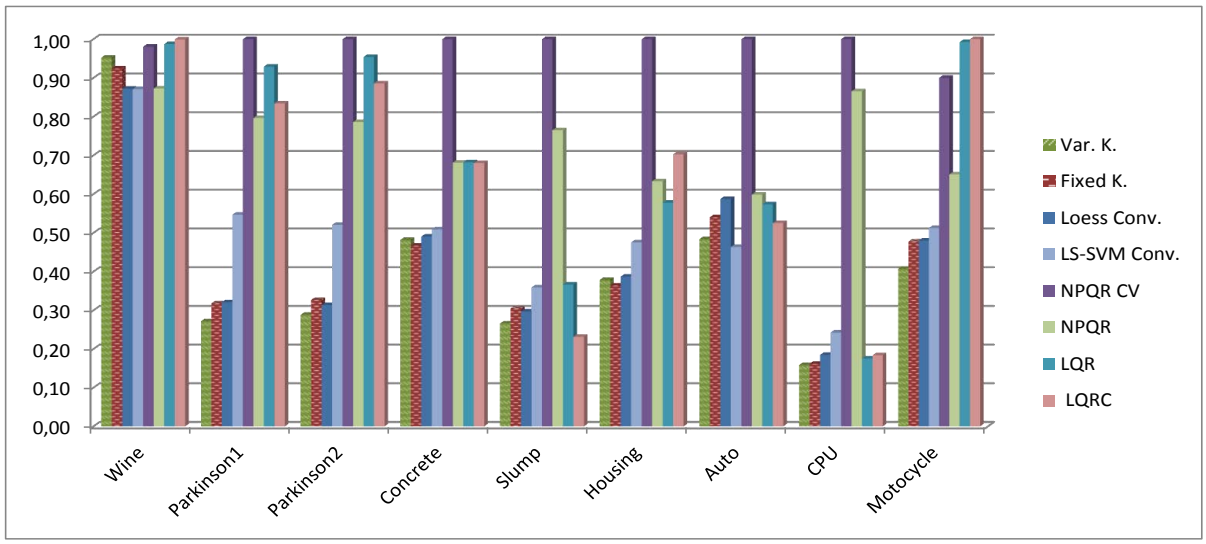

Figure 9.12: EGSD chart for benchmark datasets with $\beta = 0.99$. The smallest value denotes the most efficient band. This measure ignores the reliability.

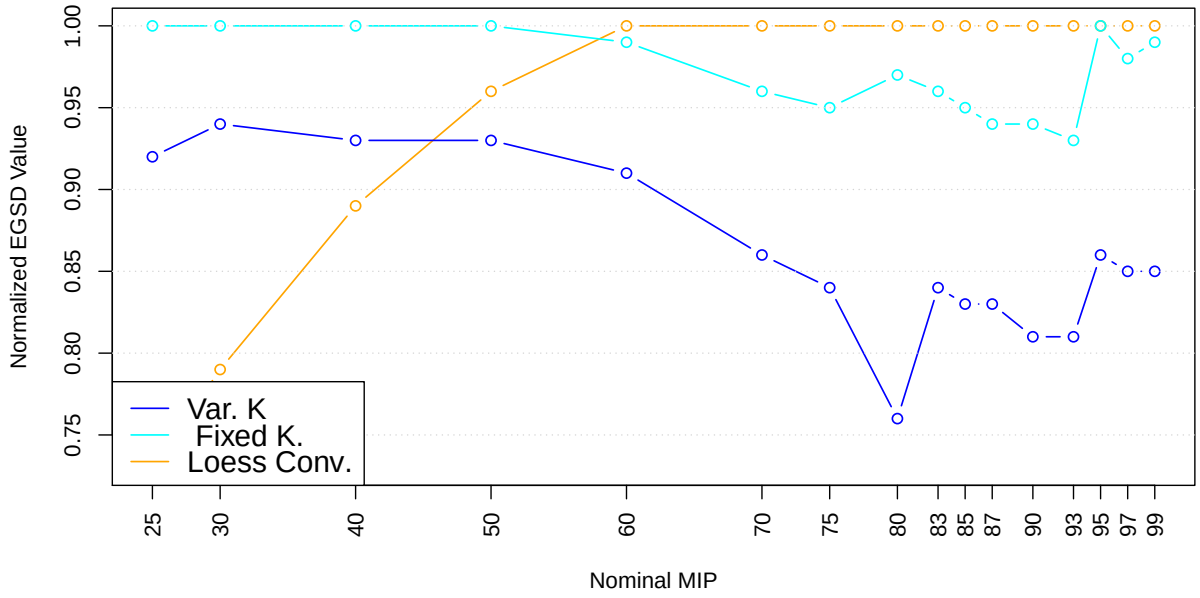

Figure 9.13: EGSD plot for Parkinson1 dataset. The lowest line denotes the method that yields the most efficient band. This measure ignores the reliability.

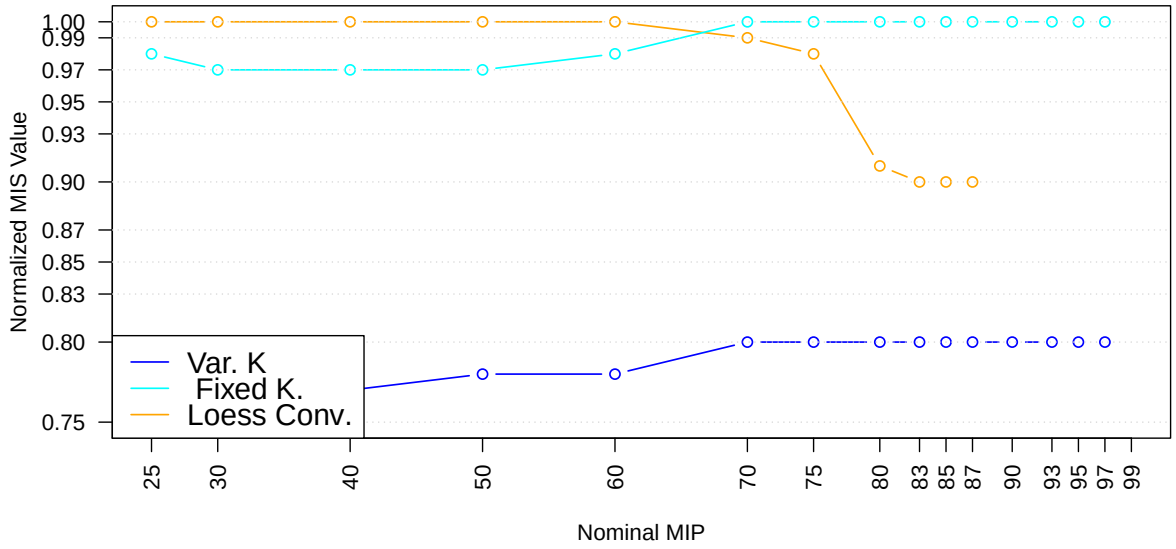

Figure 9.14: MIS plot for Parkinson1 dataset. The smallest value denotes the tightest reliable band.

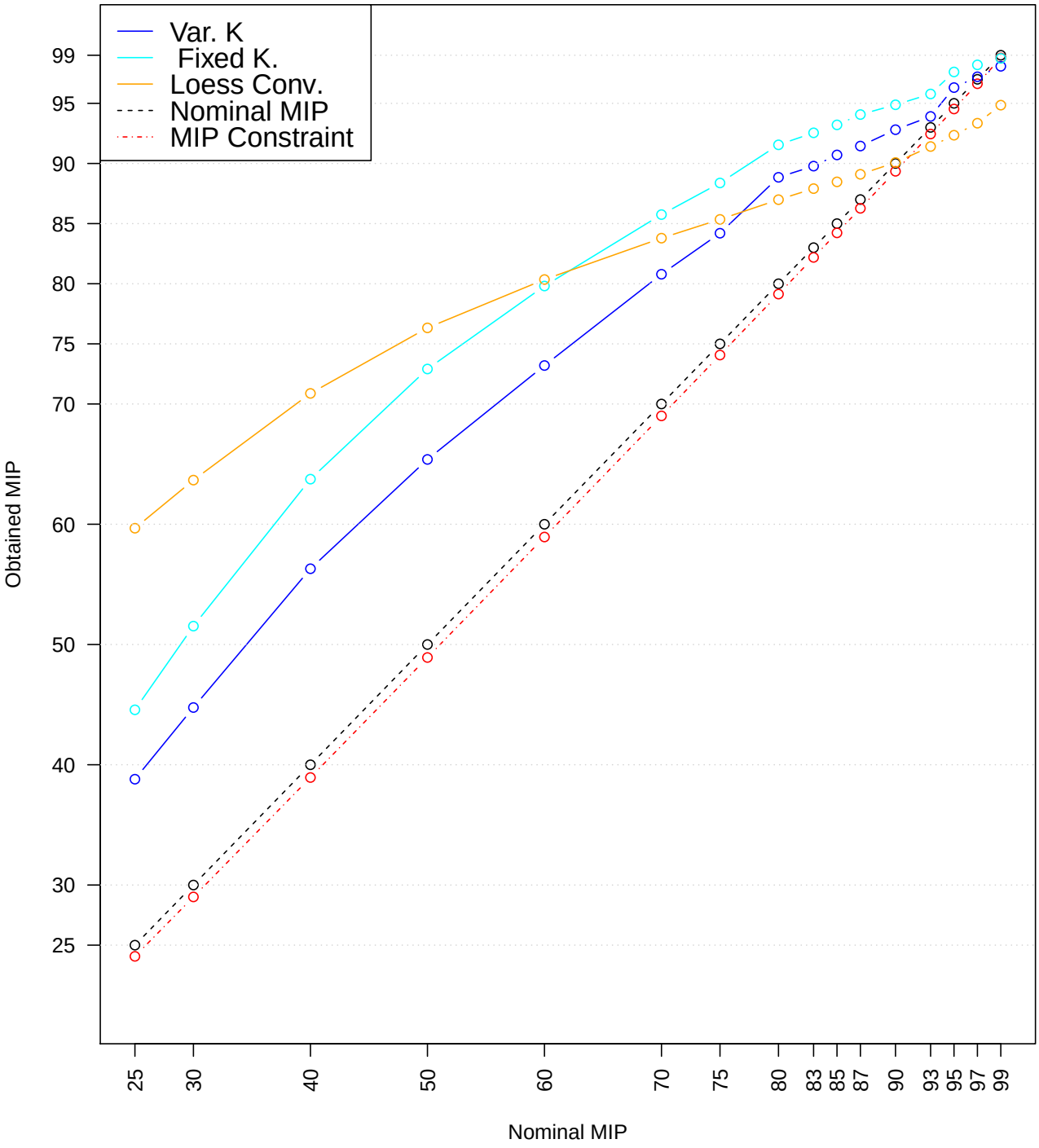

Figure 9.15: MIP plot for Parkinson1 dataset.

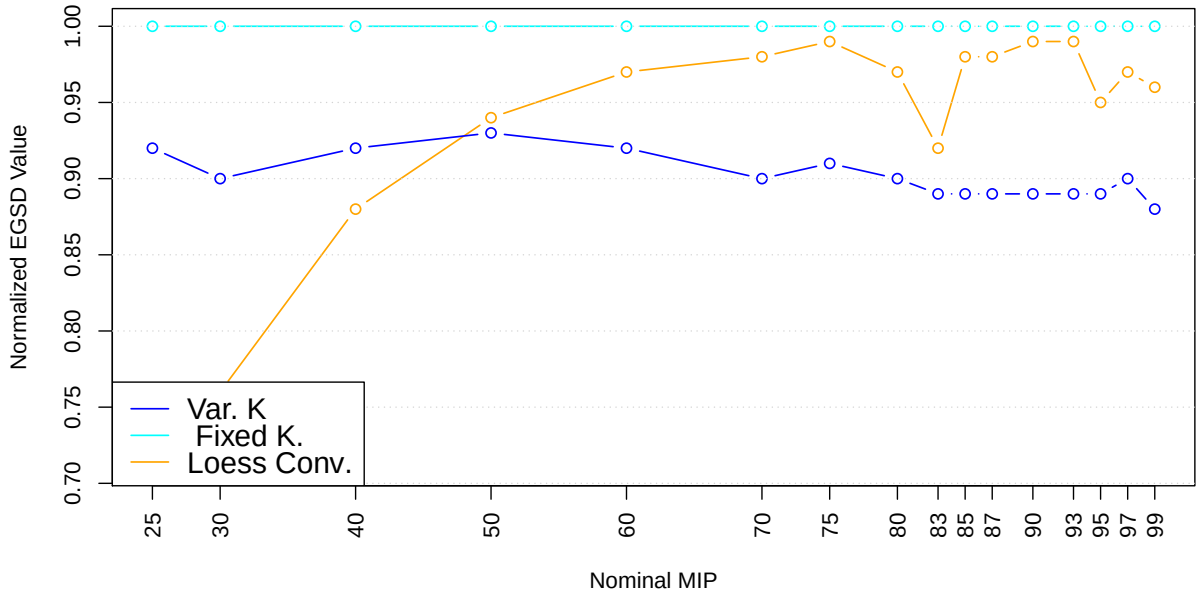

Figure 9.16: EGSD plot for Parkinson2 dataset. The lowest line denotes the method that yields the most efficient band. This measure ignores the reliability.

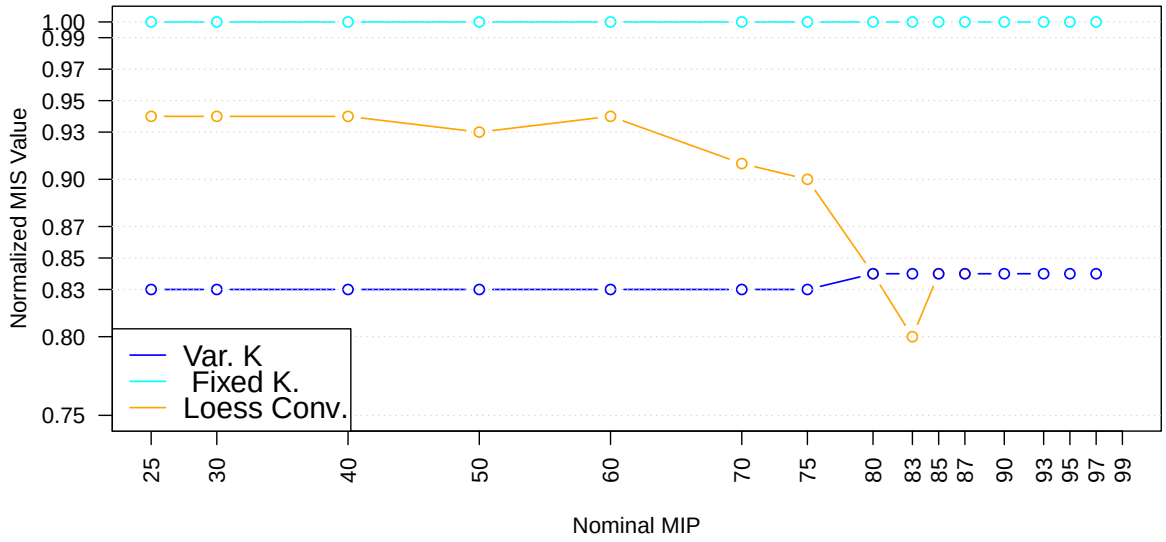

Figure 9.17: MIS plot for Parkinson2 dataset. The smallest value denotes the tightest reliable band.

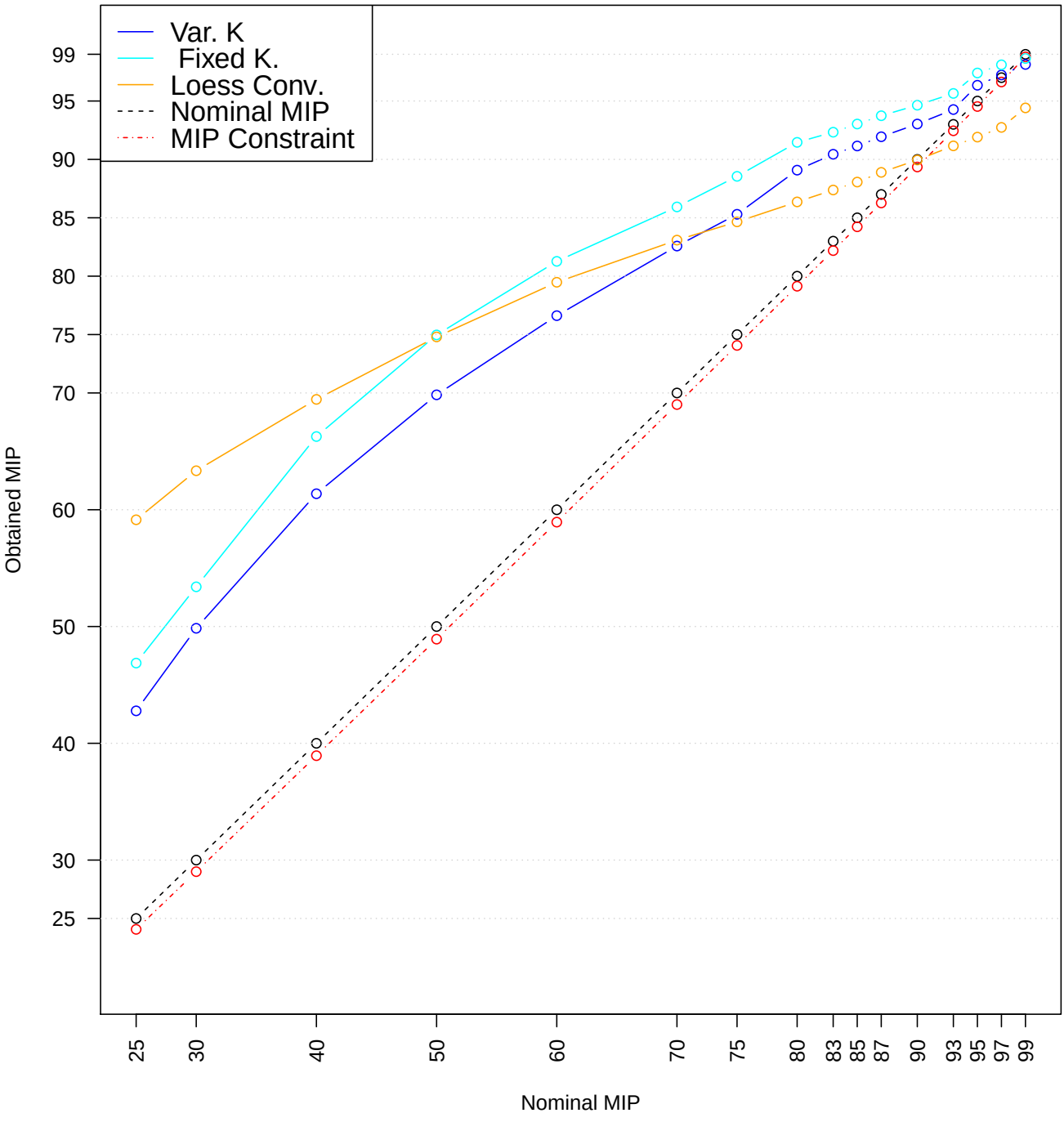

Figure 9.18: MIP plot for Parkinson2 dataset.

9.3.4 Discussion of Results

We have compared our introduced methods with six other well-known interval prediction methods. This comparison is performed with a 10-fold cross validation schema on nine benchmark regression datasets which contain between 1 and 21 predictors. For $\beta \geq 0.9$, it becomes a critical task to find an effective predictive interval method that works on all datasets. However these inter-quantiles are the most used ones in machine learning and statistical hypothesis testing. Hence, we first compared the mentioned methods based on their strengths while providing β-predictive interval models with $\beta = 0.8, 0.9, 0.95$ and 0.99. While comparing our methods with their six competitors, we found them to be the most reliable non-linear predictive interval models. Our experiments have shown that they are usually also the most effective solution.

The conventional methods "Loess Conv." and "LS-SVM Conv." are revealed to be unreliable solutions. They even fail for $\beta = 0.9$ although they are almost always less efficient than "Var K." and "Fixed K." and **their envelope is almost always larger than the "Var K." model's band**. There is just one case where "Fixed K." and "LQRC" are more reliable than "Var K.". However "LQRC" always provides much wider bands than our methods and it is also much more inefficient and imprecise than "Var K." and "Fixed K.". On the other hand, if we ignore their reliability, "Loess Conv." and "LS-SVM Conv." rank are the most efficient methods after "Var K.". They sometimes provide tighter bands than "Fixed K." , however a model which provides a tight band but usually does not work is not appropriate for predictive interval models. "NPQR CV" is more reliable than the conventional pair but it is the least efficient solution. "NPQR" and "LQR" are absolutely not appropriate for high confidence interval prediction.

In a second attempt, we compared our methods with their most effective competitor. These comparisons have been performed on the five largest datasets of the nine benchmarks and each time on 16 distinct desired contents (β value). These experiments show the superiority of "Var K." and then "Fixed K.". **We have seen that "Var K." usually provides models with the tightest bands and they are almost always the most effective and more precise than others.** "Fixed K." models are usually more effective and precise than "Loess Conv.". Note that for $\beta \geq 0.5$, "Loess Conv." is the most effective solution but it is in the same time the least precise. By more effective, we mean that the normalized EGSD value is the smallest for $\beta \geq 0.5$ but does not provide the tightest and the most precise envelope. Thus we do not recommend "Loess Conv.", because its model provides intervals that are too wide.

In a regression context, the conditional mean, the conditional variance and/or the conditional quantile may have different functions. The conditional mean is the general trend of the regression function whereas the conditional quantile is more related to the local distribution of the response variable. Least-squares based interval prediction methods ("Loess Conv.", "LS-SVM Conv.", "Fixed K." and "Var. K") try to indirectly estimate the conditional quantile function. They first estimate the conditional mean and then, based

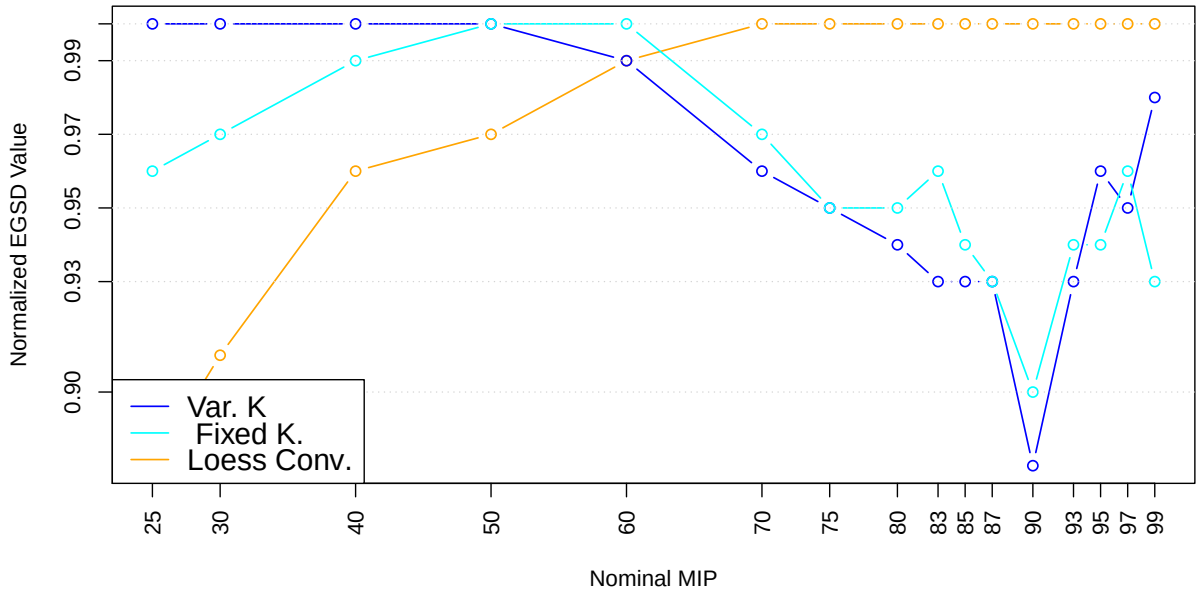

Figure 9.19: EGSD plot for Concrete dataset. The lowest line denotes the method that yields the most efficient band. This measure ignores the reliability.

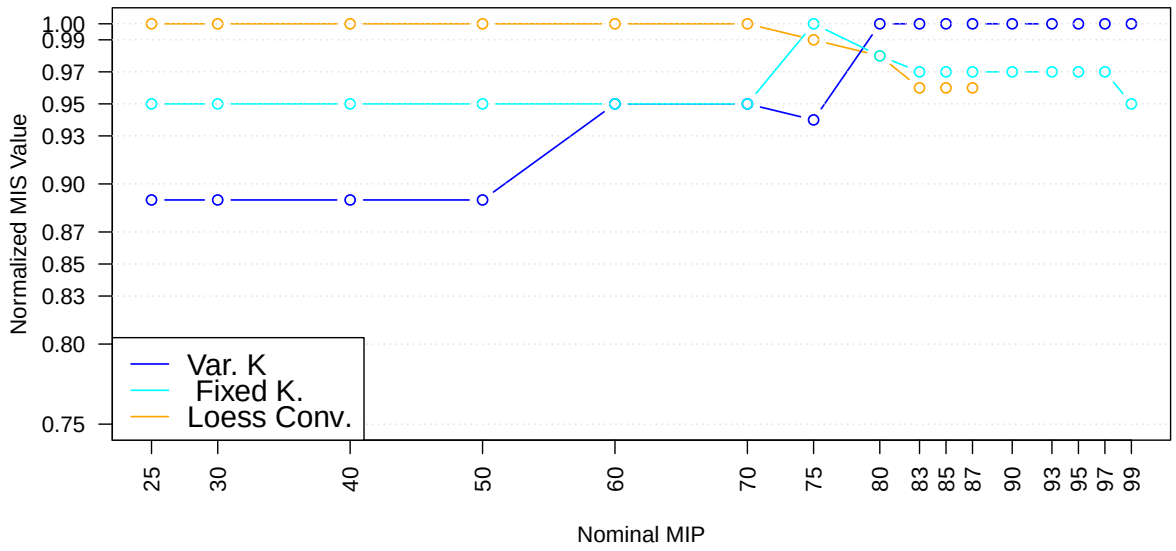

Figure 9.20: MIS plot for Concrete dataset. The smallest value denotes the tightest reliable band.

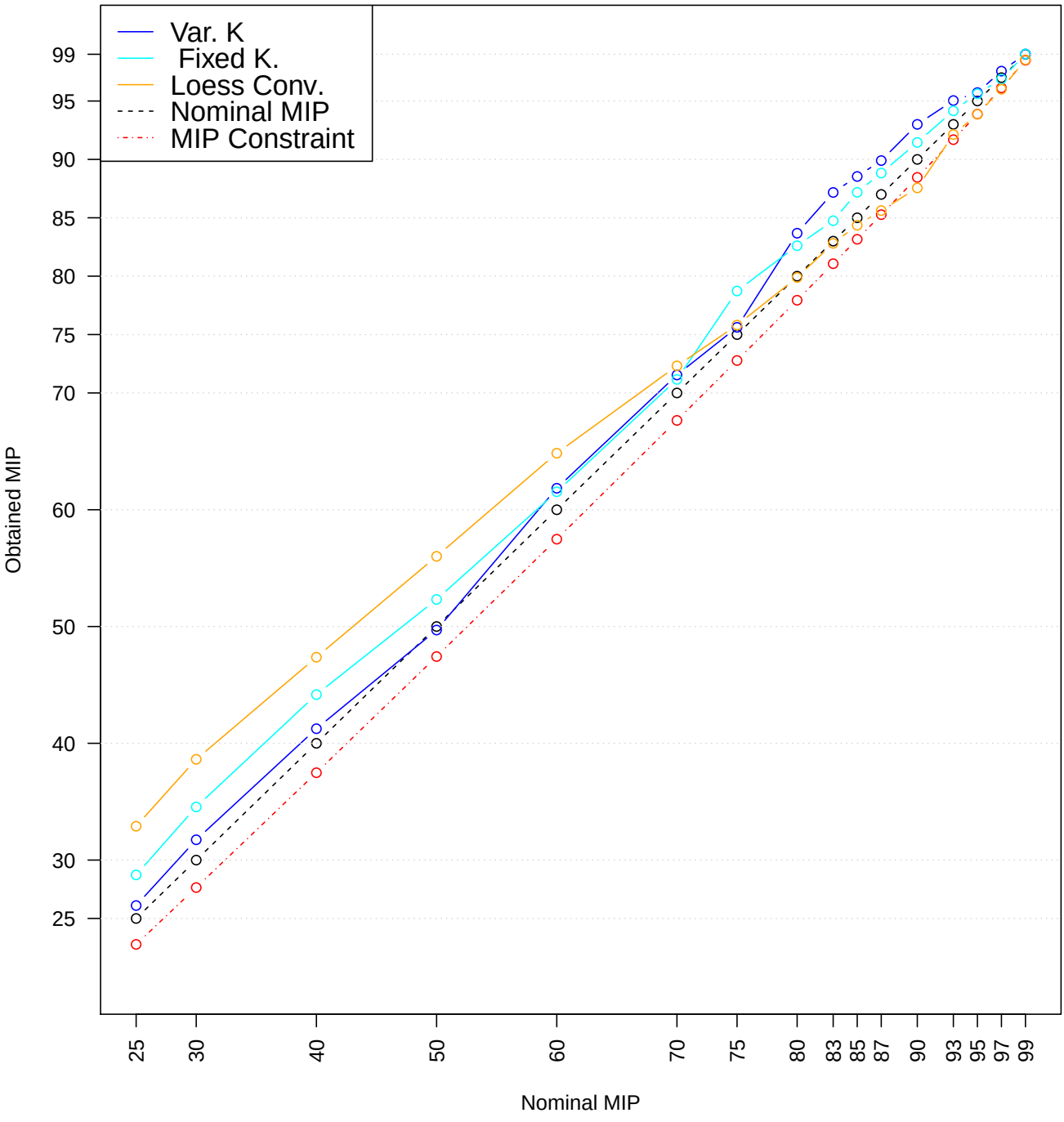

Figure 9.21: MIP plot for Concrete dataset.

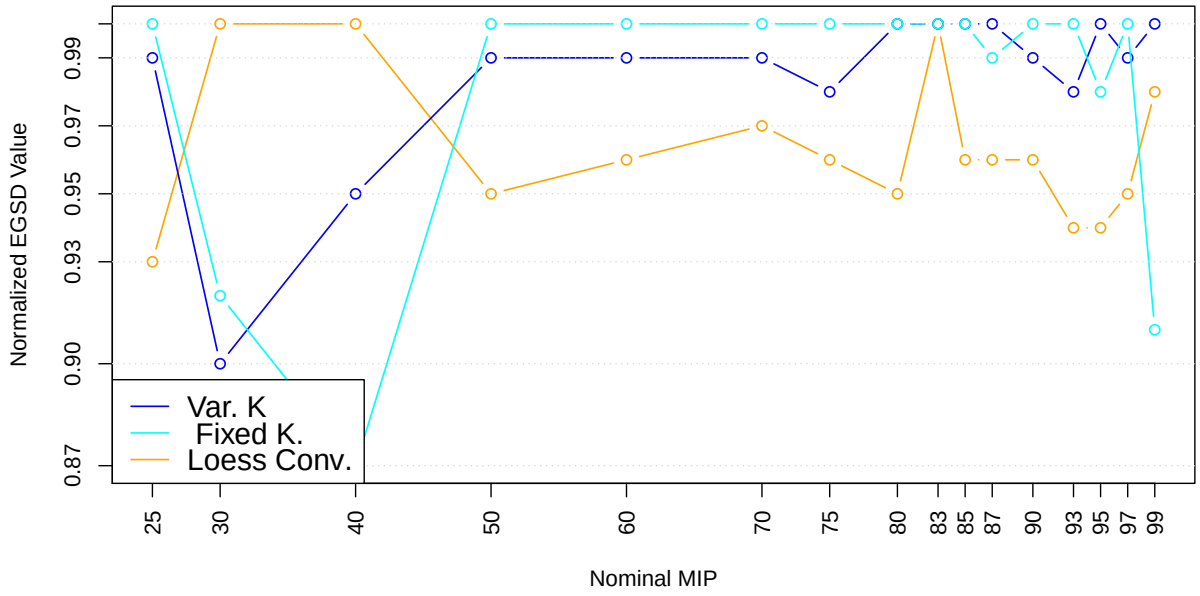

Figure 9.22: EGSD plot for Wine dataset. The lowest line denotes the method that yields the most efficient band. This measure ignores the reliability.

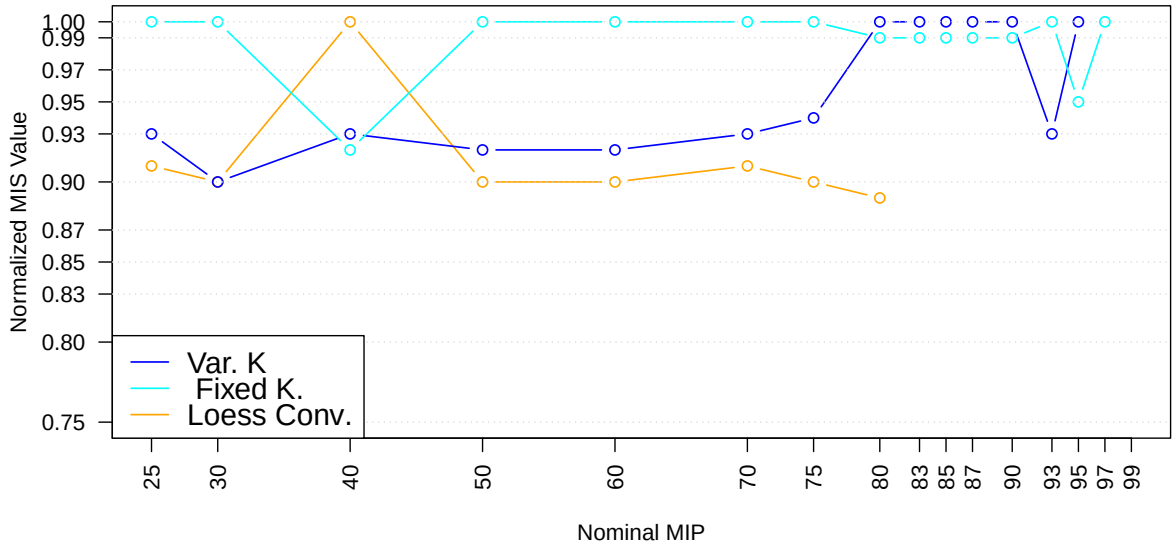

Figure 9.23: MIS plot for Wine dataset. The smallest value denotes the tightest reliable band.

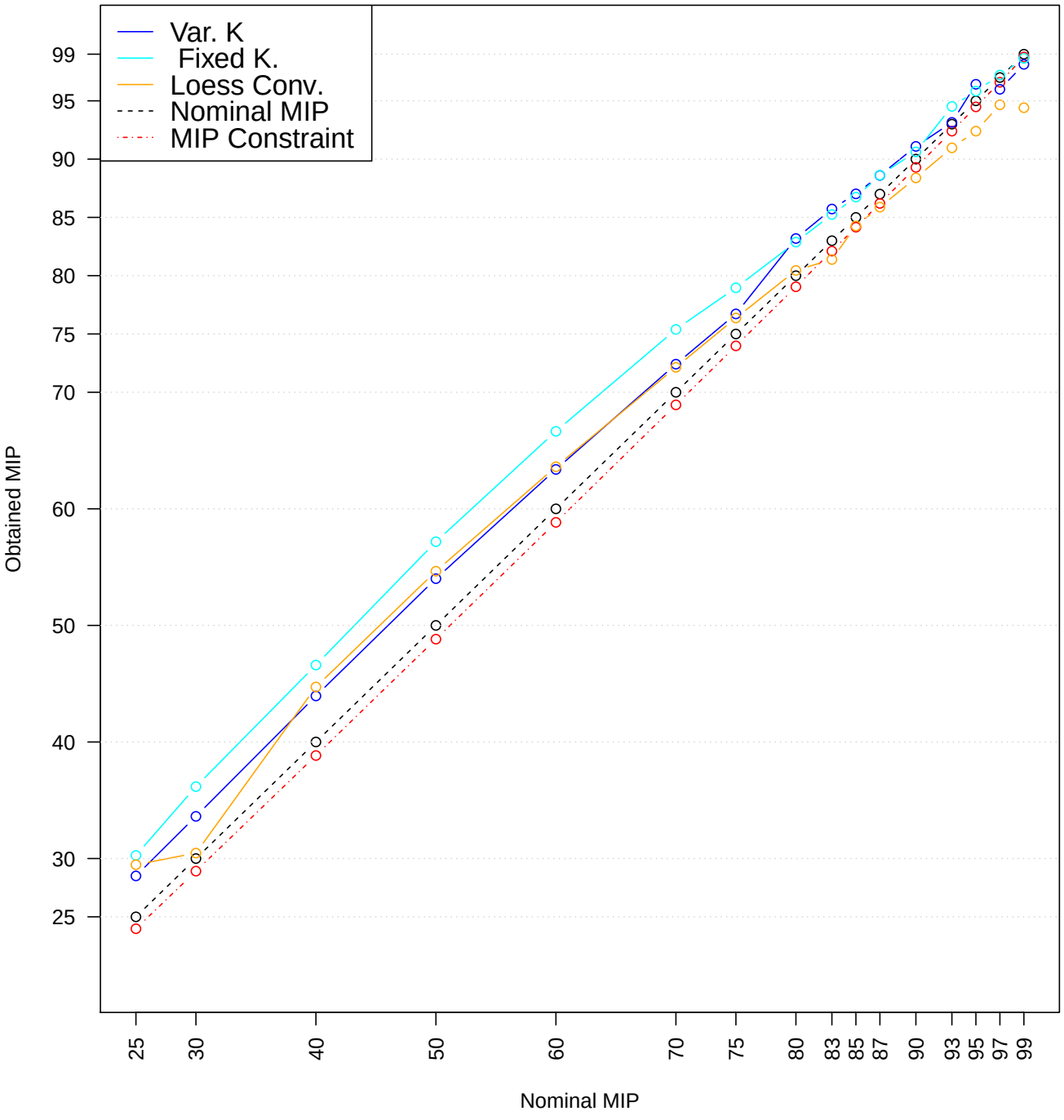

Figure 9.24: MIP plot for Wine dataset.

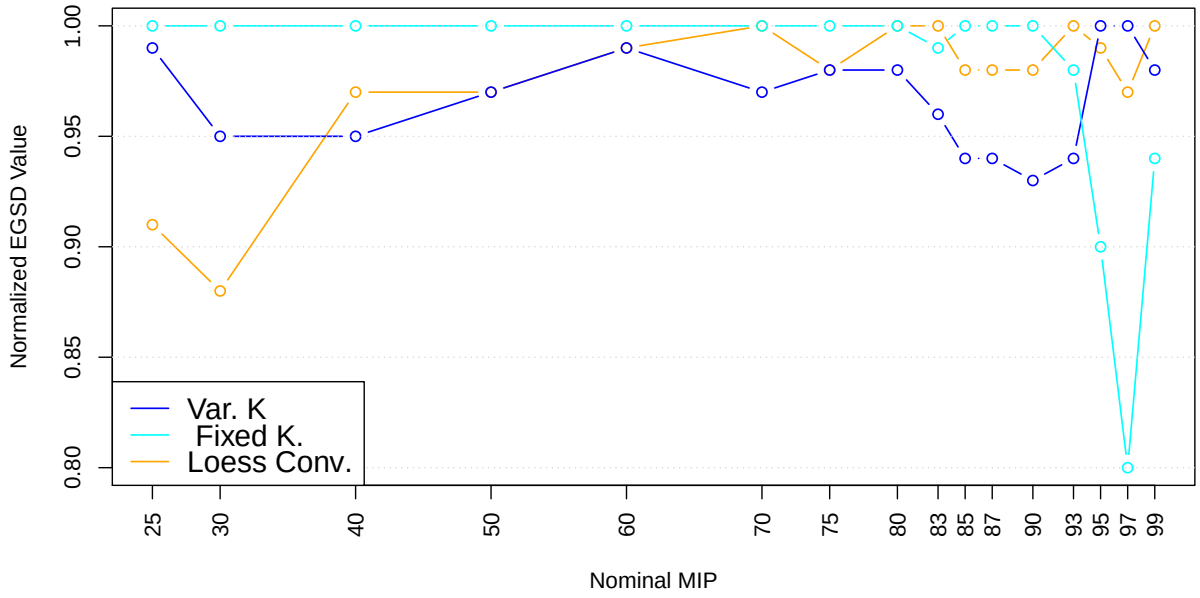

Figure 9.25: EGSD plot for Housing dataset. The lowest line denotes the method that yields the most efficient band. This measure ignores the reliability.

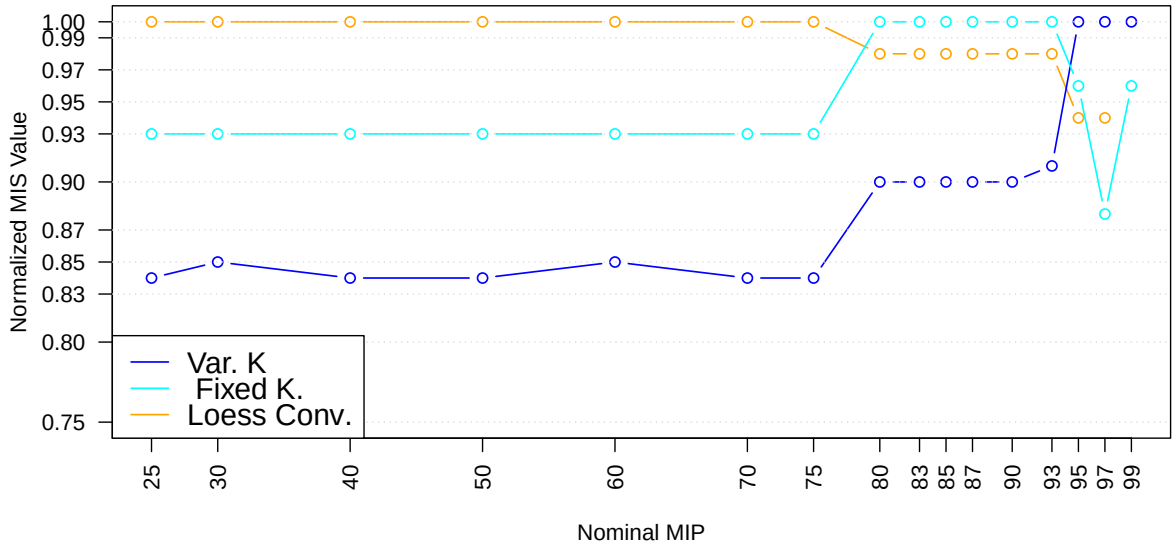

Figure 9.26: MIS plot for Housing dataset. The smallest value denotes the tightest reliable band.

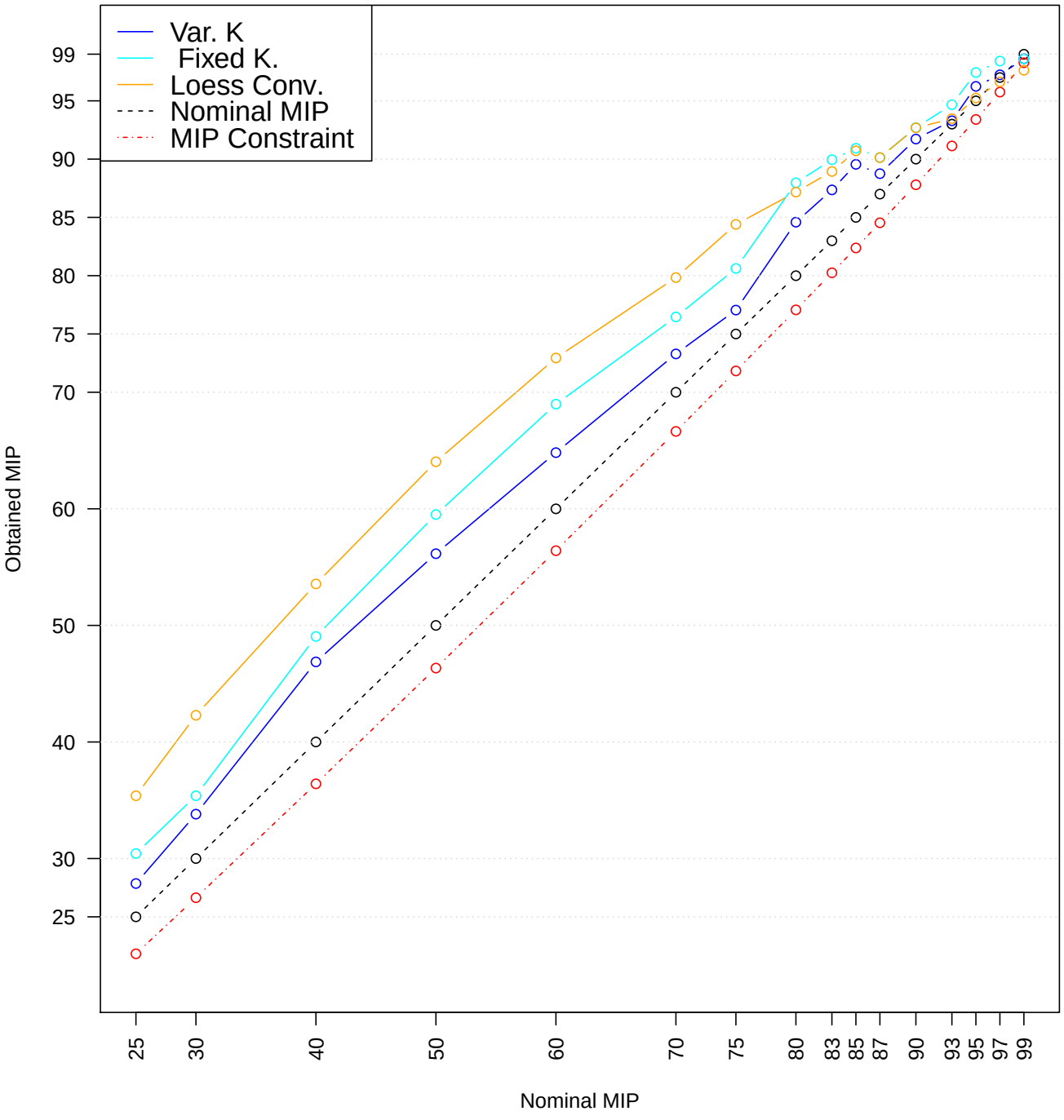

Figure 9.27: MIP plot for Housing dataset.

on this estimated conditional mean, they estimate the conditional quantile. On the other hand, we have quantile regression based methods ("NPQR", "NPQR CV", "LQRC" and "LQR") which directly estimate the conditional quantile. We know that the general trend is easier to predict and its estimator, compared to the conditional quantile, has a higher speed of convergence. This is why all of our least-squares based interval prediction methods are more efficient than the quantile regression based methods. Another reason for this superiority may be the absence of a conditional quantile function. It can occur where the conditional variance of the error distribution is not a function of the predictors. Our proposed methods are in the class of least-squares based interval prediction methods, so they take advantage of this fast convergence. However they are more reliable and efficient than the other member of this class (conventional methods). This is because our methods take into account the sample size and find confidence intervals on inter-quantile of the local distribution for the response variable whereas the conventional methods just estimate asymptotic global inter-quantiles of the conditional response variable. This idea is explained in detail in 7.5.

9.4 Experiments for Simultaneous Predictive Intervals for KNN

In this section we will describe the interval prediction methods which will be used to build simultaneous predictive interval models.

Implementation of Method

All the interval prediction methods listed below are explained in Chapter 5, except for our predictive-interval method for local linear regression, which is introduced in Chapter 7. Our selected methods as follows:

- "KNN Var. K": two-sided Simultaneous Predictive Intervals for KNN, as explained in 8.3.
- "KNN Conv." the conventional interval prediction method explained in 5.1.1, obtained with a KNN regression.

We have to mention that we use the Tricube kernel as the kernel function in all of our experiments.

Hyper-parameter tuning

In a first attempt, datasets are divided into two subsamples of size $\frac{2}{3}n$ and $\frac{1}{3}n$, where n represents the dataset size. The part containing $\frac{2}{3}$ of observations are used to tune the predictive interval model's hyper-parameters. The hyper-parameters are MIN_K, MAX_K and γ for our proposed interval regression method and just K for the KNN ("KNN Conv.").

For the classical KNN, the fixed K maximizing the Root Mean Squared Error (RMSE) of response variable is chosen. For our proposed method, the hyper-parameters having the smallest MIS and also satisfying the simultaneous MIP constraint (see (8.2)) are selected.

Experiments Plan

Once we have tuned the hyper-parameters, all the instances will serve to validate the results using a 10-cross validation schema. For each desired proportion of simultaneous inclusion (β value), we compare the reliability and the the tightness of the obtained band of the tested models. The goal is to find simultaneous β-content predictive interval models where $\beta = 0.9, 0.95$ and 0.99. The motivation of these β values is that these inter-quantiles are the most used ones in machine-learning and statistical hypothesis-testing. Another reason justifying our choice is that they are harder to approximate.

When considering the simultaneous interval regression, it is expected that the fraction of prediction values inside the envelope (for each of the 10 models in cross validation) will be greater than or equal to β (Simultaneous MIP constraint). For example, for $\beta = 0.95$ in a 10-fold cross validation, it is expected that each of the 10 built model to have a Mean Inclusion Percentage (MIP) greater than or equal to 0.95 ($MIP \geq \beta$). In our experiments, we are interested to compare the obtained intervals regardless of any variable selection or outliers detection preprocessing. The results are the mean inclusion percentages and the Mean of Interval Size (MIS) in each of the fold in the 10-fold cross validation scheme. The MIP (see (8.2)) and MIS over all the 10-fold cross validations are also contained in the results.

9.4.1 Results

Table 9.6 summarizes the application of the algorithm 4 ("KNN Var. K") and the conventional interval prediction approach combined with KNN ("KNN Conv.") to the seven datasets listed below. For each 10-fold cross validation scheme, the following quality measures are computed:

- MFIP: Mean Fold Inclusion Percentage (value of the MIP for one fold). It must be greater than or equal to the desired β for each of the 10 models built in the cross validation phase. This is the simultaneous MIP constraint explained in Equation (8.2).
- Min(MFIP): minimum value of MFIP across the 10 models. $\min(MFIP) < \beta$, *means that the underlying model failed to cover the required proportion* β *of the response values.*

The column MIS is the Mean of Interval Size for all the 10 models and σ_{is} is the standard deviation of the interval size over the whole dataset. Note that σ_{is} is not defined for the conventional method because its interval size is constant over the entire data set. The star

* sign appears when $\min(MFIP)$ satisfies the requirement (i.e. $\min(MFIP) \geq \beta$). When only one of the two compared methods satisfies this requirement, the result is given in bold.

Dataset	Algo.	90%		95%		99%	
		Min(MFIP)	$\overline{MIS}$ (σ_{is})	Min(MFIP)	$\overline{MIS}$ (σ_{is})	Min(MFIP)	$\overline{MIS}$ (σ_{is})
Parkinson1 (n=5875, p=21)	KNN Conv.	94.54 *	6.62	94.55	7.88	95.4	10.36
	KNN Var. K	90.98 *	5.01 (6.92)	**95.23 ***	6.38 (8.75)	**99.14 ***	11.19 (14.47)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (5, 40, 0.25)$		$(MIN_K, MAX_K, \gamma) = (5, 40, 0.35)$		$(MIN_K, MAX_K, \gamma) = (5, 40, 0.8)$	
Parkinson2 (n=5875, p=21)	KNN Conv.	94.04 *	4.73	94.55	5.64	95.57	7.41
	KNN Var. K	92.34 *	3.97 (5.37)	**95.23 ***	5.06 (6.77)	**99.14 ***	9.37 (11.85)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (5, 25, 0.3)$		$(MIN_K, MAX_K, \gamma) = (5, 25, 0.4)$		$(MIN_K, MAX_K, \gamma) = (5, 25, 0.87)$	
Wine (n=4898, p=12)	KNN Conv.	78.93	1.84	90.59	2.19	93.46	2.88
	KNN Var. K	**90.2 ***	2.5 (0.55)	**95.71 ***	3.51 (1.48)	98.77	5.04 (1.05)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (20, 50, 0.9)$		$(MIN_K, MAX_K, \gamma) = (5, 25, 0.99)$		$(MIN_K, MAX_K, \gamma) = (20, 50, 0.999)$	
Concrete (n=1030, p=9)	KNN Conv.	80.58	25.58	86.4	30.48	94.17	40.05
	KNN Var. K*	**91.26 ***	33.29 (11.86)	**95.14 ***	41.91 (14.8)	**99.02 ***	80.72 (26.47)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (10, 25, 0.6)$		$(MIN_K, MAX_K, \gamma) = (10, 25, 0.7)$		$(MIN_K, MAX_K, \gamma) = (10, 25, 0.99)$	
Auto (n=398, p=8)	KNN Conv.	87.17	9.96	90	11.87	94.87	15.6
	KNN Var. K	**94.87 ***	12.57 (6.48)	**95 ***	14.98 (7.72)	97.43	23.54 (11.98)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (7, 20, 0.95)$		$(MIN_K, MAX_K, \gamma) = (7, 20, 0.95)$		$(MIN_K, MAX_K, \gamma) = (7, 20, 0.99)$	
Housing (n=506, p=14)	KNN Conv.	84.31	14.23	90.19	16.96	94	22.29
	KNN Var. K	**92.15 ***	22.9 (13.09)	**96 ***	27.28 (15.6)	98	43.45 (24.44)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (10, 20, 0.99)$		$(MIN_K, MAX_K, \gamma) = (10, 20, 0.99)$		$(MIN_K, MAX_K, \gamma) = (10, 20, 0.999)$	
Slump (n=103, p=10)	KNN Conv.	80	12.73	80	15.16	80	19.93
	KNN Var. K	**90 ***	29.58 (9.83)	90	35.25 (11.71)	**100 ***	46.32 (15.4)
	Hyper. params.	$(MIN_K, MAX_K, \gamma) = (5, 15, 0.99)$		$(MIN_K, MAX_K, \gamma) = (5, 15, 0.99)$		$(MIN_K, MAX_K, \gamma) = (5, 15, 0.99)$	

Table 9.6: Comparing the interval prediction method proposed to provide simultaneous predictive intervals for KNN.

For $\beta = 0.9, 0.95$ and for the datasets Parkinson1 and Parkinson2 in Table 9.6, we can see that our method gives smaller intervals than the KNN Conv. approach. However, contrary to the KNN Conv. approach, the intervals contain the required proportion β of the response values. It is usually a difficult task to satisfy the requirement for $\beta = 0.99$ and it becomes even harder for small dataset. Because each fold contains $\frac{n}{10}$ of total observations, so one percent is equal to $\frac{n}{1000}$. It means that the constructed intervals must miss at most $\frac{n}{1000}$ of total instances and this is a quite hard task for small and even medium-sized dataset. But we can see that our method satisfies this condition for half of the datasets and the mean of the inferred intervals is compared to the required constraint. It is also interesting to note that our proposed method performs better in general for bigger datasets. This is because our method is based on the local density of the data.

9.4.2 Results Discussion

The results show that our approach performs very well on dense dataset. In the case of dataset with small sample size compared to the number of variables, our method is less reliable, but it is still better than the conventional interval prediction method in KNN.

9.5 Conclusion

This chapter investigated two concepts by experiments: the first concept was the proposed pair of non-parametric predictive interval methods which were introduced in chapter 6 and the second concept was simultaneous predictive intervals for KNN regression.

The first concept was studied in detail. We used several regression datasets to compare our predictive interval method for local linear regression with six well-known other interval prediction methods. The selected methods have been tested on their capacity to provide two-sided β-content predictive interval models. While comparing our methods with their six competitors, we found our methods to be the most reliable non-linear predictive interval models. **We have seen that "Var K." generally provides models with the tightest bands and they are almost always more effective and precise than others.** Then it comes to "Fixed K." models which are normally more effective and precise than their conventional competitors. The conventional interval prediction methods reveals to be unreliable solutions. They even fail for $\beta = 0.9$, although they are almost always less efficient than our predictive interval methods and **their envelope is almost always larger than the "Var K." model's band.** If we ignore the reliability notion, the conventional method ranks as the most efficient method after our predictive interval method. It sometimes provides tighter bands than "Fixed K.". However, a model which provides a tight band but usually does not satisfies the reliability constraint (MIP test), is not appropriate for high confidence interval prediction. Then we explained our experiments on the simultaneous predictive models with KNN regression. This part mentioned the results published in [Ghasemi Hamed 12c]. The results show that our method performs very well on large datasets. In the case of

datasets with small sample sizes compared to their number of variables, our method is less reliable, but it is still better than the conventional interval prediction method with KNN.

The next chapter discusses the ground-based aircraft trajectory prediction problem, which is a critical issue for air traffic management. A safe and efficient prediction is a prerequisite for the implementation of automated tools that detect and solve conflicts between trajectories. We modeled this problem by regression, and so we will use our proposed methods to find the tightest reliable band.

Chapter 10

Predictive Interval Models: Application to Aircraft Trajectory Prediction

Contents

Ground-based aircraft trajectory prediction is a critical issue for air traffic management. A safe and efficient prediction is a prerequisite for the implementation of automated tools that detect and solve conflicts between trajectories. This chapter has been partially published in [Ghasemi Hamed 13]. In this work, a standard point-mass model and statistical regression

method is used to predict the altitude of climbing aircraft. In addition to the standard linear regression model, two common non-linear regression methods, LS-SVM and loess are used. These methods lead to five different prediction models and they are compared based on their point based prediction performance. However because of the critical nature of our problem and regarding the safety constraints, it seems more reasonable to predict intervals rather than precise aircraft positions. So we apply nine different interval prediction methods on our aircraft trajectory prediction dataset. Some of these interval prediction models are built upon the obtained prediction models and others (Quantile regression based models) are constructed without using the preceding regression models. The experiments part compares these models based on their reliability, efficiency and the tightness of the obtained envelope.

A dataset is extracted from two months of radar and meteorological recordings, and several potential explanatory variables are computed for every sampled climb segment. A Principal Component Analysis allows us to reduce the dimensionality of the problems, using only a subset of principal components as input to the regression methods. The prediction models are scored by performing a 10-fold cross-validation. Statistical regression method results appears promising. The experimental part shows that the proposed regression models are much more efficient than the standard point-mass model. Our interval prediction models have the advantage of being more reliable and narrower than those found by other interval prediction and point-mass models. The chapter is organized as follows: the first section introduces the aircraft trajectory prediction problem. Then Section 10.2 describes the point-mass model and reviews its equations. The third section describes how the regression methods are applied to our problem and the experimental results are given in the fourth section.

10.1 The aircraft trajectory prediction problem

This section begins by describing the aircraft trajectory prediction context. We will have a quick review of the ground based trajectory prediction motivations. Next we survey the state of the art of the problem, and then we will explain and present arguments our solutions.

10.1.1 The context

Predicting aircraft trajectories with great accuracy is central to most operational concepts ([Swenson 06], [Consortium 07]) and is necessary to the automated tools that are expected to improve air traffic management (ATM) in the near future. On-board flight management systems predict the aircraft trajectory using a point-mass model of the forces applied to the center of gravity. This model is formulated as a set of differential algebraic equations that must be integrated over a time interval in order to predict the successive aircraft positions in this interval. The point-mass model requires knowledge of the aircraft state (mass, thrust, etc), atmospheric conditions (wind, temperature), and aircraft intent (target speed or climb

rate, for example).

Much of this information is not available to ground-based systems, and the available information is not known with good accuracy. The actual aircraft mass is currently not transmitted to the ATM ground systems, although this is being discussed in the EUROCAE[1] group in charge of elaborating the next standards for air-ground data-links. For a recent reference on the mass estimation problem see [Alligier 13]. The atmospheric conditions are estimated through meteorological models. Finally, the current ground-based trajectory predictors make fairly basic assumptions on the aircraft intent (see the "airlines procedures" that go with the BADA[2] model. These default "airline procedures" may not reflect reality, where the target speeds are chosen by the pilots according to a cost index that is a ratio between the cost of operation and the fuel cost. These costs are specific to each airline operator, and are not available. As a consequence, ground-based trajectory prediction is currently fairly inaccurate, compared with the on-board prediction. A simple solution would be to downlink the on-board prediction to the ground systems. However, this is not sufficient for all applications: some algorithms ([Durand 96, Swenson 06, Consortium 07, Drogoul 09, Prats 10, Chaloulos 10]) require the computation of a multitude of alternate trajectories that could not be computed and downlinked fast enough by the on-board predictor. There is a need to compute trajectory predictions in ground systems, for all traffic in a given airspace, with enough speed and accuracy to allow a safe and efficient 4D-trajectory conflict detection and resolution. The literature on trajectory prediction is fairly large, and one may refer to [Musialek 10] for a literature survey on the subject, or [Gong 04], [Romanelli 09], or [Vivona 10, Tastambekov 14] for the trajectory predictor's statistical analysis and validation. Other works focus on the benefits provided to ground-based trajectory predictors by using additional, more accurate, input data ([Center 98], [Coppenbarger 99], [ADA 09]). An interesting model-based stochastic approach is presented in [Lymperopoulos 06], although this is only validated in a simulation environment.

10.1.2 Our approach

In this work, we compare different ways of dealing with the trajectory prediction problem, focusing on the aircraft climb with a 10 minute look-ahead time. We are also interested in finding intervals which contain a desired (i.e. 0.95) proportion of the future aircraft position. Such intervals reflect the prediction uncertainty and can be used for more accurate conflict detection. Climb phase prediction has already been treated by Alligier et al. [Alligier 12]. Their work addresses the energy rate prediction problem during the climb phase. We selected the climb phase because predicting during this phase is harder and much less accurate than during the cruise phase. As a first approach, the point-mass model is applied with different settings for the model parameters, considering a constant CAS/Mach climb procedure where the aircraft first climbs at a constant Calibrated Air Speed (CAS) until it reaches the CAS/Mach crossover altitude and then continues the climb at a constant Mach

[1]EUROCAE:European Organization for Civil Aviation Equipment

[2]BADA:Base of Aircraft DAta

number. In this approach, the basic parameter setting consists of using the standard CAS and Mach values of the BADA climb procedures file, and a standard reduced thrust during climb, with an average reference aircraft mass. The second setting still uses the reference mass and standard thrust reduction factor, but the actual CAS is computed from the past aircraft positions.

The second approach is radically different and is based on regression methods. The predicted aircraft position is considered as a function $f(x)$, where x is a vector of input variables and $f(\cdot)$ is parametric or non-parametric function. We have also applied Possibilistic KNN regression to the trajectory prediction problem [Ghasemi Hamed 12a]. This consists of predicting possibility distributions rather than precise values. This method focuses on finding a conditional possibility distribution for the K-nearest neighbors (KNN) regression method.

In this work, the regression input variables are the past aircraft positions, the observed CAS at the current altitude, the deviation of the air temperature from the standard atmosphere, and the predicted wind at different flight level. The regression must be adjusted using historical data so that the computed output fits the observed position as closely as possible. In this work, we will use three well-known regression methods. The idea is to see how a parametric linear model, a common parametric non-linear model and an efficient non-parametric model perform on our dataset. In a first attempt, we use an OLS to predicts the altitude z of the aircraft based on the past trajectory. The next model is a LS-SVM regression model which belongs to the class of parametric non-linear models. The third model is the loess [Cleveland 88] method. It is a version of locally weighted linear regression which uses K-nearest neighbors as its bandwidth. For more details on loess see Chapter 4. As discussed before, aircraft trajectory prediction is a critical problem and we need more than point based prediction models. Thus, we will look for the interval prediction method that provides the smallest reliable envelope. We employed the following interval prediction methods: our predictive interval methods for LLR, the conventional interval prediction method, tolerance intervals for linear regression, and linear quantile regression.

10.2 The point-mass model

10.2.1 Simplified model

Most ground systems use a simplified point-mass model, sometimes called a *total energy model*, to predict aircraft trajectories. This model, illustrated in figure 10.1, describes the forces applying to the center of gravity of the aircraft and their influence on the aircraft acceleration, making several simplifying assumptions[3]. It is assumed that the thrust and drag vectors are colinear to the airspeed vector, and that the lift is perpendicular to these

[3]Note that more complex point-mass models have been proposed for UAV or fighter airplanes (see [T. Kinoshita 06]), modeling also the side-slip angle.

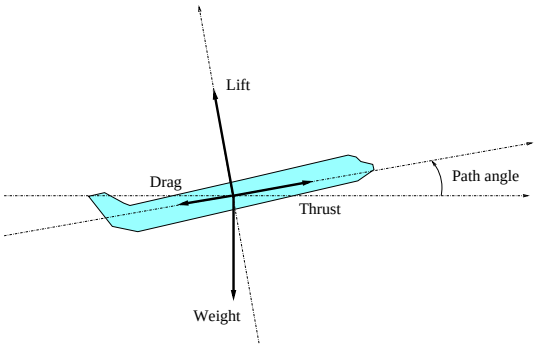

Figure 10.1: Simplified point-mass model.

vectors. Thus, projecting the forces on the airspeed vector axis, the longitudinal acceleration $a = \frac{dV_{\text{TAS}}}{dt}$ along the true airspeed (V_{TAS}) axis can be expressed as follows:

$$m.a = T - D - m.g.sin(\gamma), \tag{10.1}$$

where T is the total thrust, D the aerodynamic drag of the airframe, m the aircraft mass, g the gravitational acceleration, and γ the path angle (i.e. the angle between the airspeed vector and the horizontal plane tangent to the earth surface).

Introducing the rate of climb/descent $\frac{dh}{dt} = V_{\text{TAS}}.sin(\gamma)$, where h is the altitude in meters, this equation can be rewritten as follows (see [Nuic 09]):

$$(T - D).V_{\text{TAS}} = m.V_{\text{TAS}}.\frac{dV_{\text{TAS}}}{dt} + m.g.\frac{dh}{dt}. \tag{10.2}$$

Several equivalent forms of this equation can be used (see Eurocontrol BADA User Manual), depending on which unknown variable is being calculated from the other known variables. Actually using Equation (10.2) to predict a trajectory requires a model of the aerodynamic drag for any airframe flying at a given speed through the air. In addition, we may need the maximum climb thrust, which depends on the engines that the aircraft is equipped with. In the experiments presented here, the Eurocontrol BADA model was used for that purpose.

One cannot use (10.2) without prior knowledge of the initial state (mass, position, speed,...) of the aircraft, and also of the pilot's intentions as to how the aircraft will be operated in the future (thrust law, speed law, or rate of climb). When the aircraft is operated at a given calibrated air speed (CAS[4]) or Mach number, computing the true air speed (TAS) requires knowledge of the atmospheric conditions (the air temperature and pressure). Finally, as we need to predict the trajectory over the ground surface, and not only through the air, the wind magnitude and direction are also required.

[4]CAS: Calibrated Air Speed, which can be equated to the speed indicated on the pilot's intruments.

10.2.2 Aircraft operation during climb

Generally, when no external constraint applies during the climb, the aircraft is operated at constant CAS and variable Mach number, until a specified Mach number is reached. Above this CAS/Mach crossover altitude, the aircraft is operated at a constant Mach number, and variable CAS. External constraints may apply, however. After take-off, the aircraft cannot exceed a specified maximum CAS until Flight Level is reached100[5]. This first climb segment is followed by an acceleration at FL100, and then a second climb segment at a higher calibrated air-speed, until the CAS/Mach crossover altitude is reached.

In this work, we consider only this second climb segment at constant CAS, followed by the constant Mach climb, as we are mostly interested in predicting the aircraft trajectory in the en-route airspace. Note that some other air traffic control constraints may apply, that modify the aircraft operation during climb. For instance, the aircraft may be operated at a prescribed rate of climb, on some flight segments, in order to be above a specified flight level over a given waypoint.

Even without such constraints, and assuming a climb at constant CAS/Mach, predicting the aircraft trajectory is not easy for ground systems. The actual CAS and Mach values are chosen by the airlines' operators, according to a cost index specific to each airline. The cost index, and the chosen CAS and Mach values are not known by the air traffic control systems, although some studies show the improvements that such knowledge would provide in the trajectory prediction ([Center 98],[Coppenbarger 99]).

10.3 The Aircraft trajectory Prediction dataset

In our trajectory prediction problem, we predict the altitude $z(t)$ at time $t > t_0$, where t_0 is the current time, The input x is a vector of values extracted from the values:

- The current and previous aircraft states, characterized by $z[k]$, $d[k]$, with $k \in [-10, 0]$. The past trajectory is sampled every δt seconds. $z[k]$ denotes the value measured for the altitude z at time $t = t_0 + k\delta t$. With this notation, $z[0] = z(t_0)$ is the current altitude, $z[-1]$ is the altitude δt seconds before t_0, and so on. The same notation applies for the distance d;
- The difference between the actual air temperature at sea level and the air temperature of the International Standard Atmosphere (ISA) at sea level;
- The along-track and cross-track wind w and the temperature T at different altitudes;
- Other variables, such as the current *CAS*, *Mach number*, *energy share factor*, *ROCD*, *Ground speed*, etc, and their derivatives with respect to time.

The regression model must be adjusted using historical data, so that the computed outputs are as close as possible to the observed data. The performance of the tuned model

[5]FL100 = 10000 feet above isobar 1013 hPa.

is measured by assessing how the model generalizes on fresh inputs. K-fold cross validation can be used for that purpose. In order to start with a relatively simple problem, we predict only one future point of the trajectory, N steps ahead. Let us now describe the dataset used to predict the future aircraft position.

10.3.1 The available data

Recorded radar tracks from the Paris Air Traffic Control Center were used to build the patterns used in the regression methods. This raw data is made of one position report every 1 to 3 seconds, over two months (July 2006, and January 2007). In addition, the wind and temperature data from Meteo France are available at various isobar altitudes over the same two months. The raw Mode C altitude[6] has a granularity of 100 feet. So the recorded aircraft trajectories were smoothed, using a local quadratic model, in order to obtain: the aircraft position (X,Y in a projection plan, or latitude and longitude in WGS84), the ground velocity vector (V_x, V_y), the smoothed altitude (z, in feet above isobar 1013.25 hPa), and the rate of climb or descent (ROCD). The wind (W_x, W_y) and temperature (T) at every trajectory point were interpolated from the meteo datagrid. The temperature at isobar 1000 hPa was also extracted for each point, in order to compute a close approximation of $(\Delta T_0)_{\text{ISA}}$, the difference between the actual temperature and the ISA model temperature at isobar 1013.25 hPa (mean sea level in the ISA atmospheric model). This $(\Delta T_0)_{\text{ISA}}$ is one of the key parameters in the BADA model equations.

Using the position, velocity and wind data, we computed the true air speed (TAS), the distance flown in the air (dAIR), the drift angle, and the along-track and cross-track winds (W_{along} and W_{cross}). The successive velocity vectors allowed us to compute the trajectory curvature at each point. The actual aircraft bank angle was then derived from true airspeed and the curvature of the air trajectory. The climb, cruise, and descent segments were identified, using triggers on the rate of climb or descent to detect the transitions between two segments.

Finally, the BADA model equations were used to compute additional data, such as: calibrated airspeed (CAS), Mach number (M), energy share factor[7] (ESF), as well as the derivatives of these quantities with respect to time.

10.3.2 Filtering and sampling climb segments

As our aim is to compare several prediction models, we focused on a single aircraft type (Airbus A320), and selected all flights of this type departing from Paris Orly (LFPO) or Paris Roissy-Charles de Gaulle (LFPG). We selected the Airbus A320, because this the most common aircraft in Europe. Another technical reason is that introducing other aircraft types forces us to treat the aircraft trajectory prediction problem with more complex models

[6]This altitude is directly derived from the air pressure measured by the aircraft. It is the height in feet above isobar 1013.25 hPa.

[7]The energy share factor (ESF) measures how much of the energy is devoted to climb or to longitudinal acceleration.

having more parameters and requiring significantly more trajectories. In fact, if we are able to obtain efficient prediction for a single aircraft type, then the investigation of a more complex model, which is a much bigger task, could be easily justified. The trajectories were then filtered so as to keep only the climb segments. An additional 40-seconds were trimmed from the beginning and end of each segment, so as to remove climb/cruise or cruise/climb transitions. The trajectories were then sampled every 15 seconds, with time and distance origins at the point P_0 where the climb segment crosses flight level FL180[8]. The trajectory segments were sampled so as to obtain 10 points preceding P_0, and a number of points after P_0, depending on the chosen look-ahead time. So the trajectory observed during the preceding time steps (2 minutes 30 seconds), can be used to predict the aircraft position at one or several future time steps. The predicted position can be compared to the actual aircraft position at the same time step.

Trajectories exhibiting a bank angle greater than 5 degrees were discarded, so that the influence of trajectory turns on the rate of climb can be neglected. This allows us to disregard the lateral navigation in our trajectory prediction problem, and focus on the longitudinal and vertical dimensions of the trajectory.

10.3.3 Construction of the regression dataset

The regression models $y = f(x)$ are tuned and assessed using sets of patterns (x, y_d), where x is an input vector, and y_d is the corresponding desired output that can be compared to the computed output y. These patterns, that we have already described in Section 10.3, were extracted from the sampled climb segments. 1500 patterns were randomly chosen, to build the set used in our experiments.

Each pattern used for regression contains the current ground speed, true and calibrated air speed, Mach number, and their derivatives with respect to time, the energy share factor, the altitude variations and distance flown for the ten preceding time steps, and also the predicted wind and temperature at several altitudes that the aircraft may cross in the look-ahead time. It also contains the potential target variables: distance flown, in the air or above the ground, and altitude reached after N time steps in the future.

10.3.4 Principal component analysis

The final patterns set contains 79 numerical variables, measured for 1500 aircraft climbs. There are 76 explanatory variables, and 3 variables to explain (although only the altitude is predicted here). A principal component analysis was performed on the explanatory variables, so as to reduce the dimensonality and avoid redundant input variables in the trajectory prediction. Figure 10.2 shows the standard deviations of the principal components: 9 components have a standard deviation above 1, and 7 other components are between 0.5 and 1.

[8]FL180: 18000 feet above isobar 1013 hPa.

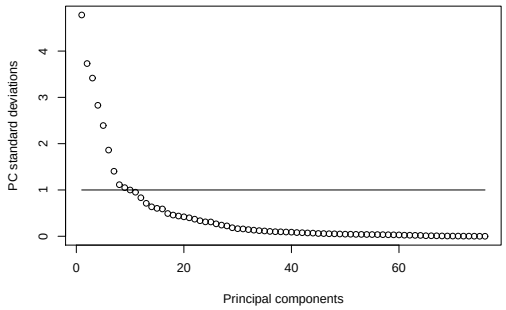

Figure 10.2: Principal components standard deviations.

Principal components are linear combinations of the initial variables, that we can use as explanatory variables in the regression method. This reduces the dimensionality from 10 to 15 significant principal components, instead of the 76 initial variables. One must keep aware, however, that using linear combinations representing projections on a basis of orthogonal vectors may not take into account some non-linearities in the initial variables.

10.3.5 Validation of regression assumptions

As seen in Chapter 4, a regression model must have non-autocorrelated errors. So we tested our linear model with the two-sided Durbin-Watson test. The alternative hypotheses of the Durbin-Watson test is that the true model's first-order autocorrelation is not 0. Then we tested our linear model with the Breusch-Godfrey test for serial autocorrelation of errors with order up to 12. Its null hypothesis is that there is no serial correlation of the regression model residuals of any order up to 12. These results are indicated in Table 10.1 (we also tested our model with Breusch-Godfrey test separately for Lag 1, Lag 1 and 2, until Lag 1, ..., 12 and the P-value were approximately the same). The next step was dedicated to check against homoscedasticity. For this purpose we used the Breusch-Pagan test; the alternative hypothesis of which is heteroscedasticity. The normality of errors was tested as follows: externally studentized residuals of the linear regression model having a normal homoscedastic error follow a Student distribution with $n - d - 2$ degrees of freedom, where n and d are respectively the number of residuals and the dimension of the predictor space. In our case, we have a Student distribution with 1484 degree of freedom, which can be very well approximated by the standard normal distribution. So we tested the externally studentized residuals with the Shapiro-Wilk normality test and the normality assumption was rejected with a p-value of $8.466e - 08$.

Test name	Null hypotheses	P-value	Accept the null hypothesis
Durbin-Watson	residuals first order auto-correlation equal to zero	0.2989	Accept
Breusch-Godfrey	all correlation of residual with any order up to 12 equal to zero	0.1508	Accept
Breusch-Pagan	homoscedasticity of residuals	9.778e-05	Reject
Shapiro-Wilk normality test	Normality of residuals	8.466e-08	Reject

Table 10.1: Test results for the linear regerssion model on the ATM dataset.

Note that the normality test assumes that the linear model has normal and homoscedastic error but we saw that we do not have a homoscedastic error. Thus even in case of normal errors, the normality test would have failed. Table 10.1 shows that we have a regression model which does not respect the assumptions of an OLS model. The mean of the squared residuals for our linear model was 952, but it decreased to 699 for a loess regression model. The non-linear models seem to be more promising for our ATM dataset, so in the search for a better dataset fit, we will try loess and LS-SVM regression models.

10.4 Experiments

This section describes our experiments. First, we employ our point-based aircraft trajectory prediction methods. They are composed of two BADA and three regression methods. These five prediction methods lead to five distinct prediction models which are compared based on their point-based prediction performance. However, we cannot use an illusory precise point-based prediction as the future aircraft position. Moreover, we are more interested in having a interval prediction model rather than a point based prediction. Therefore, we will apply nine interval prediction methods to our dataset. These models are compared in Table 10.3.

The tested methods are scored using a 10-fold cross-validation on the Air Traffic Management (ATM) dataset described in section 10.3.3. The cross validation procedure splits this set into ten subsets. It use nine of the subsets to build the prediction model, and keeps the remaining subset to assess the model performance. This operation is repeated 10 times, cycling through the subsets. The model's performance is assessed over the ten runs, considering the mean score, the standard deviation, and also confidence intervals for the computed output.

10.4.1 Point based prediction models

We used the following aircraft trajectory prediction methods:

- **BADA** : BADA point-mass model, using the reference mass for each aircraft and BADA values for the constant CAS/Mach, assuming reduced climb (eq. 3.8.1 and 3.8.2, p.22 in [Nuic 09]), and taking account of the $(\Delta T_0)_{\mathrm{ISA}}$ temperature difference;
- **BADA(obs)**: Same BADA model as above, but using the CAS observed at t_0, and the BADA target Mach number;
- **LR** : Ordinary least squares linear regression with the *lm* function in *R*;
- **LS-SVM** : Regression with least-squares support vector machines with the *ksvm* function in the *R*'s *kernlab* package. We use *ksvm* with the following arguments: kernel="rbfdot", for using a radial basis kernel function. We also set kpar= list(sigma= 0.01) and tau = 0.01, reduced = TRUE, tol = 0.0001;
- **Loess** : linear loess with $K_{loess} = 500$ as its bandwidth and Tricube as its kernel function; for more detail see Chapter 4.

Table 10.2 shows the prediction errors (mean absolute error, and root mean squared error) over the 10 runs of the cross-validation, for all tested methods. The 15 principal components of higher variance were used as input to the regression methods. This selection was made by prior trials, successively adding the principal components until no significant improvement was observed.

Method	MAE	RMSE
BADA	1440 (79)	1824 (95)
BADA(obs)	1440 (77)	1819 (86)
LR	744 (55)	962 (72)
LS-SVM	729 (57)	952 (73)
Loess	700 (54)	908 (72)

Table 10.2: Average prediction errors (and standard deviations) on the altitude (in feet) for Airbus A320 aircraft, using 15 principal components as input, with the reference point at FL180 and a 10-minutes look-ahead time.

All regression methods perform significantly better than the BADA point-mass model. There are several factors explaining the poor performance of the point-mass models. The parameter's choice assumed a constant CAS/Mach climb at economic thrust, and the same reference mass for all aircraft, which is not actually the case in reality. Also, the regression methods use the past trajectory to predict the future altitude, whereas our BADA models do not. Using the observed CAS instead of the BADA standard CAS does not improve the results on altitude prediction.

It came as a surprise that the LS-SVM method did not really perform better than the ordinary least squares linear regression. There may be several explanations for this. Using the principal components as inputs does favor linear methods. In addition, tuning the parameters with the ordinary least squares linear regression can be done with an exact method, whereas LS-SVM methods require iterative approximations or a stochastic selection process, that may have difficulties to find the optimum when using input variables that are not very efficient in explaining the target variable(s). In fact LS-SVM, being non-linear estimators, have less bias but higher variance than OLS. Due to the high dimensionality of our dataset and the relatively small number of observations, the LS-SVM estimations suffer from high variance. This latter leads the LS-SVM's MSE to be greater than that for OLS. For more detail on the bias-variance trade-off and model complexity in regression see [Rao 99].

As expected by the theoretical properties of loess, reviewed briefly above, we can observe that this method gives the best results on our data. We can see in Table 10.2 that loess is more efficient than LS-SVM and OLS. We used a two-sided Mann-Whitney test (paired Wilcoxon signed rank test) to compare the 10 Root Mean of Squared Error (RMSE) resulted by (the 10-fold cross validation of) loess and LS-SVM. The test rejected the null hypothesis of loess RMSE's mean being greater than or equal to the LS-SVM RMSE's mean with a p-vale of 0.004883.

10.4.2 Interval prediction models

Uncertainty on the prediction can be assessed in the following ways: once the model parameters have been tuned on the training set, we can compute a theoretical 95%-confidence interval using the root mean square error (RMSE) observed on this training set, assuming a Gaussian distribution of the error in altitude. This method is the conventional interval prediction method explained in 5.1.1 and it is applied upon BADA, BADA(obs), loess and LS-SVM models. Our regression interval prediction methods are listed below:

- "Fixed K": two-sided predictive interval for linear loess as explained in 7.1 with the fixed K tolerance neighborhood with the loess regression model obtained before. The hyper-parameter for this model are $(K, \gamma) = (270, 0.99)$.
- "Var. K": two-sided predictive interval for linear loess as explained in 7.1 with the variable K tolerance neighborhood with the loess regression model obtained before. The hyper-parameters for this model are $(MIN_K, MAX_K, \gamma) = (80, 170, 0.99)$
- "LQR": two-sided interval prediction with linear quantile regression [Koenker 05]. We used the *rq* and *rq.predict* function in *R*'s *quanterg* package.
- "LQRC" two-sided Bonferroni 0.95-level confidence β-content interval obtained with two different quantile regression models as explained in "Confidence based point-wise inference" of 5.3.3. We used the *rq* and *rq.predict* function in *R*'s *quanterg* package.

We use *predict* with the following arguments: interval="confidence", type="percentile", se="boot", bsmethod= "wild".

- "Loess Conv." the conventional interval prediction method explained in 5.1.1 obtained with the loess regression model obtained before.
- "LS-SVM Conv." the conventional interval prediction method explained in 5.1.1 obtained with the LS-SVM regression model obtained before.
- "LR Tolerance." Two-side 0.95-coverage 0.95-content tolerance interval for linear regression as explained in 5.2.3.

Method	Percentage in theoretical 95% interval (MIP)	Mean Interval size of 95% interval	Predictive Interval Model ($MIP \geq 94.07$)	EGSD
BADA	92 (2.5)	6558	X	1872.98
BADA(obs)	93 (2.1)	6738	X	1859.36
LS-SVM Conv	94.19 (2.22)	3767		993.96
Loess Conv	94.79 (2.2)	3684		948.36
Fixed K	94.26 (2.33)	3606		947.77
Var K	94.33 (2.76)	**3602**		**946.15**
LR Tolerance	99.93 (0.21)	7714		1137.90
LQR	93.73 (1.96)	3837	X	1030.72
LQRC	96.86 (2.26)	4424		1027.90

Table 10.3: Different Interval prediction models for the altitude prediction (Airbus A320), with a reference point at FL180 and a 10-minute look-ahead time.

The interval prediction results are shown in Table 10.3. The second column shows the percentage of predictions, computed with instances from the validation set, that actually fall within the 95% predicted interval (MIP measure defined in Chapter 6). The third column shows the mean interval size of this obtained interval (MIS measure defined in Chapter 6). The fourth column indicates whether the model is a predictive interval model (passes the MIP test described in Chapter 6). Any model which is not an predictive interval model is distinguished by the "X" in its fourth column. The final column displays the Equivalent Gaussian Standard Deviation (EGSD) measure which is an interval prediction model's efficiency measure. The model having the smallest EGSD value is the most efficient model, for more details see Chapter 6. Table 10.3 is also divided into four vertical parts. The first part contains the BADA results. The second part describes the conventional interval prediction method applied upon the neural network regression and the loess regression models. The third part describes the results of our proposed methods and the final part shows different linear interval prediction methods.

We can observe that point-mass models give much wider intervals than the regression models. Loess based interval prediction models and particularly "Var K" and "Fixed K" provide reliable predictive interval models being much more tighter than the BADA model. We can also observe that our proposed interval prediction methods ("Var K" and "Fixed K") have the smallest reliable envelope and they are also the most efficient interval prediction models. Then we have the conventional interval prediction models ("Loess Conv" and "LS-SVM Conv") and next linear models which are still more efficient and reliable than BADA models. It is important to note that these intervals are computed on the climb phase which is really hard to estimate.

Table 10.3 shows that our proposed predictive interval methods provide the most effective models. Now, our goal is to compare in a very detailed manner the precision, reliability, efficiency and envelope tightness of our models with their most efficient competitors. If we compare our introduced methods with the most reliable method, we have to select "Linear Tolerance". However "Linear Tolerance" is considerably less efficient and it only begins to be useful for $\beta \geq 0.99$. We have seen that "Loess Conv." gives the most effective solutions after "Var K." and "Fixed K". For this purpose we will use EGSD plots and MIP plots (described in 6.4.3) to compare the performance of "Var K.", "Fixed K" and "Loess Conv." on our dataset. For each dataset, the EGSD plot compares the efficiency of the tested models and the method having the highest line in this plot is the least efficient one. MIS plot compares the envelope width of reliable models. The method represented by the bottom line provides the most reliable envelope in the MIS plot. Note that in the MIS plotted each model is plot until its failure MIP (described in 6.4.3). Once we have compared methods based on their envelope size and their efficiency, the MIP plot will help us to compare their precision and the reliability of interval prediction models. The best model in this plot is that which is represented by the nearest line to the upper side of the "Nominal MIP line". For more explanations on these plots see 6.4.3.

By looking at Figure 10.3, we can see that when the nominal MIP is greater than or equal to 0.85, the "Loess Conv." model loses its efficiency and Figure 10.4 shows that the "Var K." model is the tightest reliable model. In the same time, Figure 10.5 compares their failure MIP and precision. We can see that they have the same failure MIP 0.97, but "Var K." and "Fixed K" are more precise models than "Loess Conv.".

10.5 Conclusion

In this chapter, we have applied several methods to the prediction of altitude. The aim was to compare these methods when predicting the altitude of climbing aircraft 10 minutes ahead, starting from an initial point at flight level FL180, and possibly using the past trajectory to improve the prediction. Radar and Meteo data recorded over two months (July 2006, January 2007) was used to build a dataset of explanatory and target variables. A principal component analysis of this data allowed us to reduce the dimensionality from 10 to 15 significant components, instead of the 76 initial explanatory variables. The models are compared by performing a 10-fold cross-validation on a set of 1500 climb segments. Our

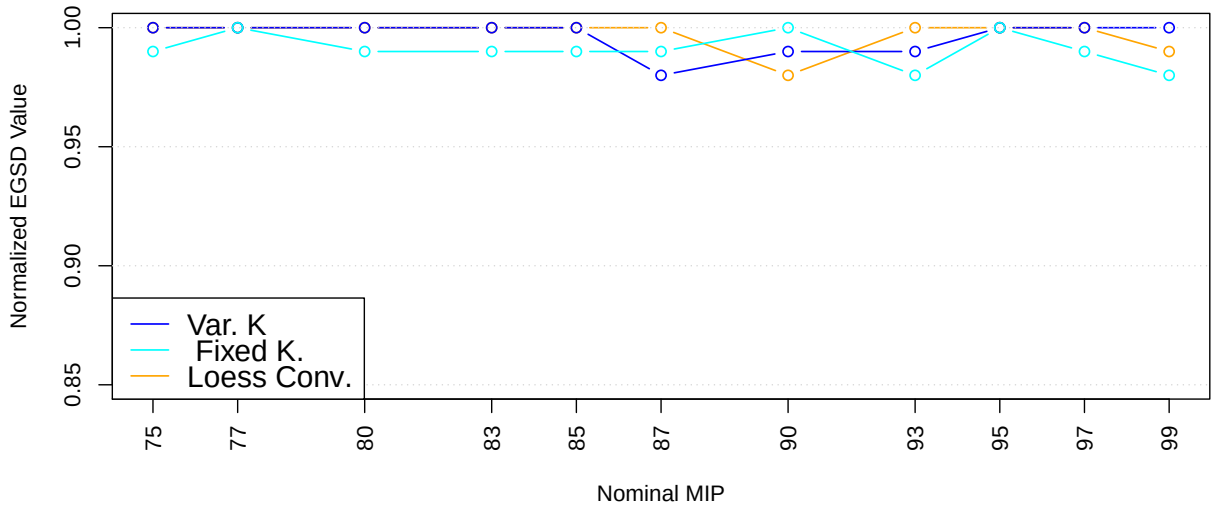

Figure 10.3: EGSD plot for the aircraft trajectory prediction datasets. The lowest line denotes method that yields most efficient band. This measure ignores the reliability.

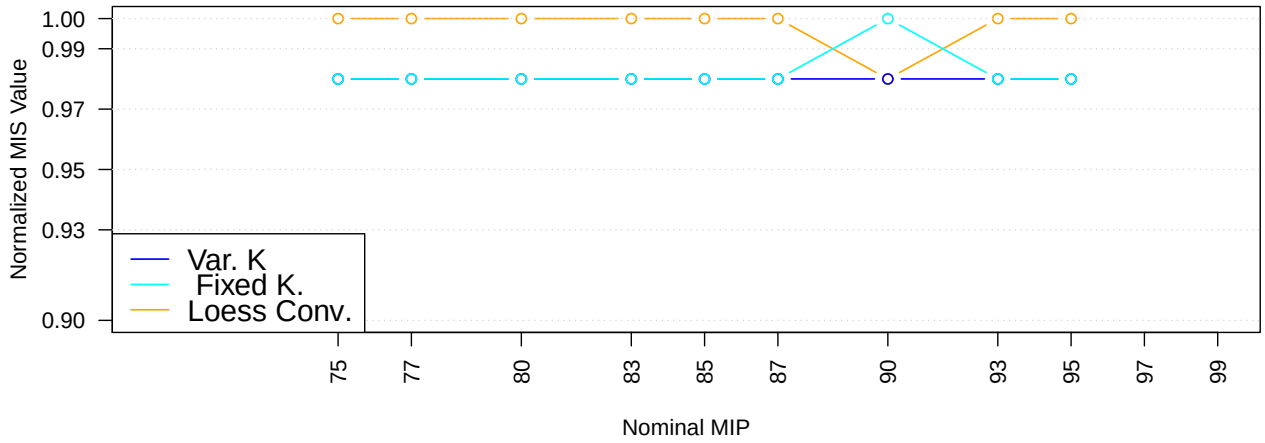

Figure 10.4: MIS plot for the aircraft trajectory prediction dataset. The lowest value denotes the tightest reliable band.

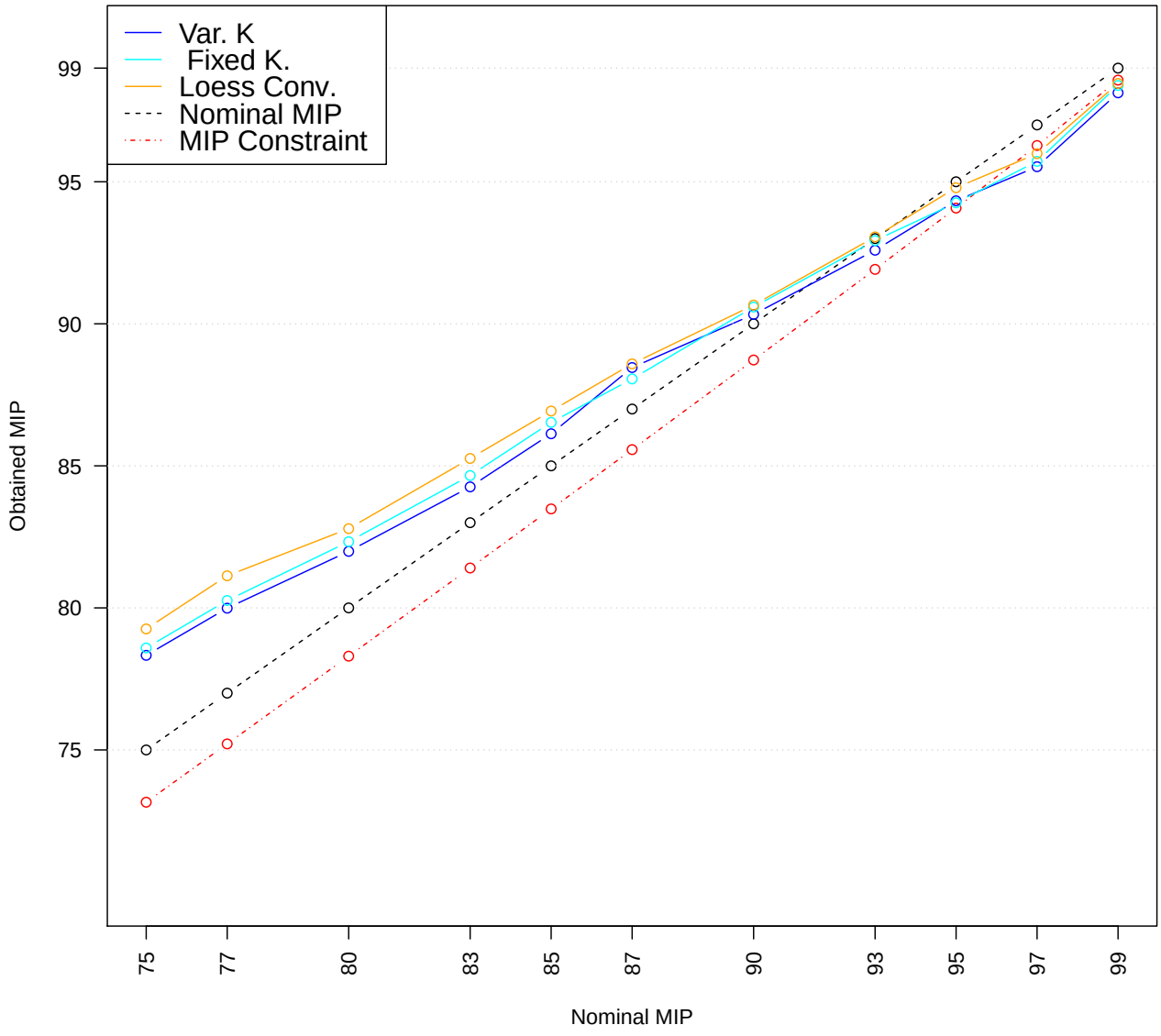

Figure 10.5: MIP plot for the aircraft trajectory prediction dataset.

results show that the regression methods perform significantly better than the point-mass model. This is not surprising as the former learns from the observation of the past trajectory, whereas the point-mass model uses the same standard values for most parameters (mass, power reduction, target speeds) for all aircraft. The linear regression method is efficient, although not as efficient as the loess regression method.

From an operational point of view, the proposed methods could be applied to the detection of potential conflicts between trajectories. The current approach of the controller is to isolate the entire vertical path of a climbing aircraft. Suppose that an aircraft (A) wants to climb until 10000 feet, so the controller will not let other aircraft (B) enter the segment 0 to 10000 feet of the climbing aircraft (A). Consequently, our interval prediction methods could be used to provide a relatively narrow probabilistic interval allowing us to detect conflicts with a great look-ahead time. In future work, we shall try to improve the loess approach by introducing elements of the point-mass model in the predictors, and by testing other robust methods. Since the regression models had efficient results, we can use bigger datasets and random effect regression models to have a production level model working with different aircraft types, destinations and trajectory prediction phases. Another parallel plan could be to conduct a more thorough analysis of the available data, to obtain less noisy data. We could then learn the aircraft trajectory in a specific operation mode, thus giving a better chance to the point-mass model.

Conclusion

This thesis has focused on high confidence two-sided interval prediction methods and their application to aircraft trajectory prediction. Our contributions are based on the classical frequentist probability framework. However we did not restrict them to aleatory model, so we also proposed a possibilistic representation of our statistical models. We began the first chapter with a review of the uncertainty frameworks that address both aleatory and epistemic uncertainty within the regression context. We explained that quantitative possibility distribution can also be viewed as a family of probability distributions. This possibility distribution contains all the probability distributions that are respectively upper and lower bounded by the possibility and the necessity measure [Didier 06]. Then we described different types of statistical intervals in the frequentist statistics and we proposed possibility distributions that encode these distinct kind of intervals. Possibility distributions encoding tolerance intervals and prediction intervals are also new concepts that we introduced in this work.

Then we extended our work in the regression context. We provided a review of interval prediction methods in the fixed regression design of the frequentist statistic. One common category contains methods that consist in building a least squares or mean estimating regression model and then employ some kind of statistical inference technique to predict such intervals. Another classical method is based on quantile regression models. We reviewed different types of intervals and described their frequentist interpretation. We also took advantage of this work to address a common interval prediction method in the Machine Learning community. However, most practitioners of this community usually employ this conventional inference for predicting interval such as prediction intervals, tolerance intervals and simultaneous tolerance intervals. We dedicated the first section of Chapter 5, to explain this conventional technique and its drawbacks. We introduced the following notions: predictive interval concept, predictive interval model, a predictive interval model test and two interval prediction measures. Predictive intervals of a linear model can be obtained with tolerance intervals for regression and confidence interval on quantile regression but they can provide wide intervals. So we explained how to tune the confidence level of tolerance intervals for regression and confidence interval on quantile regression in order to obtain efficient and reliable predictive interval models. Next, we introduced predictive interval models for local linear regression. These models provide intervals which contain at least a desired proportion of the conditional distribution of the response variable given a specified

combination of predictors. They can be obtained with tolerance intervals for regression or confidence intervals for regression quantiles but these concepts have only been treated for linear models so far. The originality of this work is to extend this concept to local linear models. Our method does not neglect the regression bias and finds intervals that work properly with biased regression models. We have also seen that all these methods can also be used for possibilistic regression with crisp input and output data.

Our predictive interval models are based on local linear regression. We assume that the mean regression function is locally linear and the prediction error variable $(Y_i - \hat{f}(x_i))$ has, locally, almost the same distribution. The idea behind this method is to exploit the local density of prediction error in the neighborhood of the query point x^* to find the most appropriate intervals that contain the desired proportion of response values $Y(x^*)$. For this purpose, we use tolerance intervals on prediction errors. They are obtained with a fixed and variable neighborhood method. We use the leave-one-out or 10-fold cross validation errors of the regression function to obtain the predictive intervals. These errors are obtained based on a local linear estimation which could be done by any bandwidth selection technique. Once the prediction errors have been found, we can use them to obtain our non-parametric predictive intervals. For this purpose, we need a second bandwidth, which is the tolerance interval bandwidth. The LHNPE bandwidth is always included in the regression bandwidth. One must not confuse our predictive intervals with bandwidth selection methods for local polynomial regression. Local linear regression needs a bandwidth, but is not just a bandwidth selection method. In the same way, our predictive interval methods are interval prediction methods which require a bandwidth on the local prediction errors.

Figure 10.5 displays the positions of our methods compared to the state of the art.

Our method differs from conventional least-squares approaches to find confidence intervals on the unknown conditional mean function because it takes into account the sample size and finds confidence intervals on inter-quantiles of the local distribution for the response variable $f(x) + \varepsilon$, while the conventional methods just estimate asymptotic global inter-quantile for the conditional response variable (or the conditional mean estimate). Most practitioners of the Machine Learning community usually estimate such predictive intervals by another conventional interval prediction method. We have seen that this method has a very small area of application and does not take into account the sample size. In the experimental part, we observed that it is one of the most unreliable predictive intervals techniques. Contrary to quantile regression, our method is based on the local linear least squares model, so one can obtain both the conditional mean function and the predictive intervals. Another main difference is that quantile regression gives estimations of quantiles which on average, finds the true quantile function but our method proposes predictive intervals which contain at least a desired proportion of the conditional response variable. Quantile regression may sometimes be more robust than least-squares estimators but it suffers from several problems. One of these problems is the absence of a conditional quantile function. It can occur if the conditional variance of the error distribution is not a function of predictors. Consider the case when the conditional quantile function is different from the

conditional mean function. We know that the conditional mean estimator converges faster than the conditional quantile estimator [Koenker 05]. Thus estimating intervals by quantile regression may be less efficient than using least-squares methods. Besides, it is important to note that quantile regression also suffers from the crossing quantile problem which is not present here. Our proposed methods are in the class of least-squares based interval prediction methods, so they take advantage of their fast convergence. However they are more reliable and efficient than the other members of this class (conventional methods). This is because our methods take into account the sample size and find confidence intervals on inter-quantiles of the local distribution for the response variable whereas the conventional methods just estimate asymptotic global inter-quantiles of the conditional response variable.

The experimental part tests our proposed method to find "predictive intervals models" and "simultaneous predictive intervals models". In this chapter, our proposed predictive interval models, other conventional interval prediction methods, linear quantile regression, confidence intervals on linear regression quantiles and a non-linear quantile regression method are applied on nine different benchmark regression datasets. The results show that our approach performs very well. It is significantly more effective than other methods and remains the most reliable non-linear interval prediction method. The advantages and drawbacks of our methods compared to the other ones are listed below:

Advantages of our approach

- It is the most reliable interval prediction method for non-linear least squares models in the experiments.
- It takes into account the amount of data available in the neighborhood to find the best trade-off between quantity of information and precision of the prediction.
- It does not ignore the non-parametric regression bias.
- It can be used with model having heteroscedastic errors.
- It directly addresses the problem of having predictive intervals that contain at least the desired proportion of response values. It is not designed to work asymptotically and also works with small datasets.
- It does not suffer from the crossing quantiles effect.
- It provides one model for two-sided interval prediction.
- It is simple, reliable and effective.
- It is based on local linear regression, which is a well-known regression method.

Drawbacks of our approach

- It is currently just based on local linear regression.
- It has a greater computational complexity than conventional and quantile regression interval prediction methods.

We explained our experiments on the simultaneous predictive models with KNN regression. This part mentioned our results published in [Ghasemi Hamed 12c]. The results show that our method performs very well on large datasets. In the case of datasets with small sample sizes compared to the number of variables, our method is less reliable, but it is still better than the conventional interval prediction method with KNN.

Finally, we have applied several methods to the aircraft trajectory problem. The aim was to compare these methods when predicting the altitude of climbing aircraft 10 minutes ahead, starting from an initial point at flight level FL180, and possibly using the past trajectory to improve the prediction. Radar and Meteo data recorded over two months (July 2006, January 2007) were used to build a dataset of explanatory and target variables. A principal component analysis of this data allowed us to reduce the dimensionality from 10 to 15 significant components, instead of the 76 initial explanatory variables. The models were compared by performing a 10-fold cross-validation on a set of 1500 climb segments. Our results show that the regression methods perform significantly better than the point-mass model. This is not surprising as the former learns from the observation of the past trajectory, whereas the point-mass model uses the same standard values for most parameters (mass, power reduction, target speeds) for all aircraft. The linear regression method is efficient, although not as efficient as the Loess regression method. From an operational point of view, the proposed methods could be applied to the detection of potential conflicts between trajectories. Our interval prediction methods could be used to provide a relatively narrow probabilistic interval allowing us to detect conflicts with a big look-ahead time.

Future Work

For future work in imprecise probability, we propose the use of tolerance intervals instead of confidence bands to infer p-boxes from statistical data. In case of predictive intervals, we have several horizons: The easiest and most promising idea is the extension of our two-sided predictive intervals to one-sided predictive intervals where they can be directly compared with confidence intervals on regression quantiles. One can also extend the predictive interval models on local linear regression to predictive interval models on any regression function, i.e. support vector machines. Interval prediction in time series models may be the next application of our methods.

The aircraft trajectory prediction may be improved by using bigger datasets with more trajectories and more aircraft types. In this case, we may better model this problem by random effect regression models. Another parallel plan could be to conduct a more thorough

analysis of the available data, to obtain less noisy data. We could then learn the aircraft trajectory in a specific operation mode, thus giving a better chance to the point-mass model. This prediction problem could also be studied by other statistical prediction models such as times series and stochastic process model and so on.

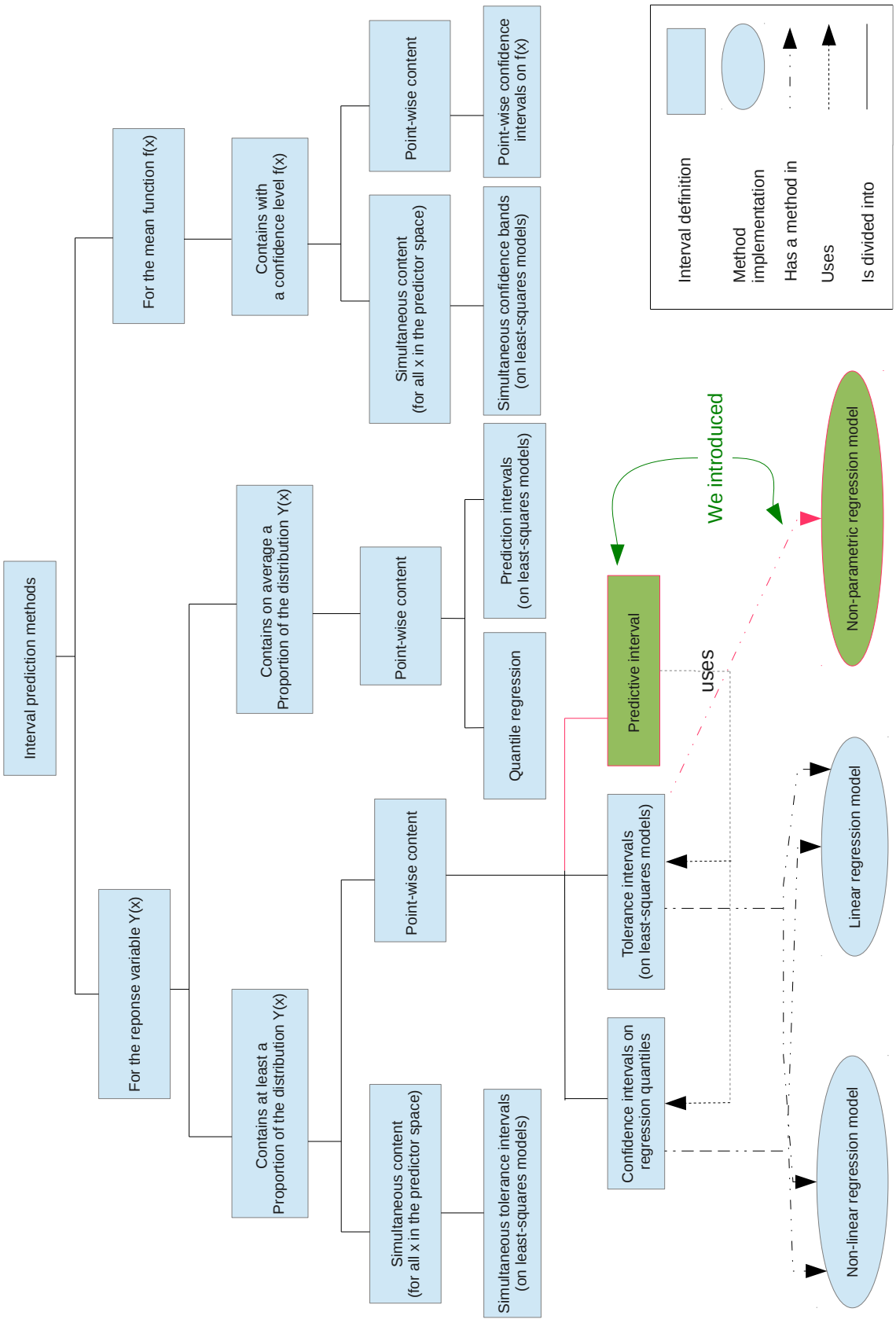

Figure 10.6: The position of the predictive intervals in the state of the art.

Glossary

ATM Air Traffic Management.

bba basic belief assignment.

BLUE Best Linear Unbiased Estimator.

cdf cumulative distribution function.

CTP distribution Confidence Tolerance Possibility distribution.

DFCTP distribution Distribution Free Confidence Tolerance Possibility Distribution.

EGSD Equivalent Gaussian Standard Deviation.

iid independent and identically distributed.

KNN K-Nearest Neighbors.

LLR Local Linear Regression.

LOO Leave-One-Out.

LPR Local Polynomial Regression.

LS-SVM Least Squares Support Vector Machines.

MSE Mean Squared Error.

OLS Ordinary Least Squares.

RMSE Root Mean of Squared Error.

SCI Simultaneous Confidence Intervals.

SDA Symbolic Data Analysis.

SVM Support Vector Machines.

TBM Transferable Belief Model.

WLS Weighted Least Squares.

References

[ADA 09] *ADAPT2. Aircraft data aiming at predicting the trajectory. Data analysis report.* Technical report, EUROCONTROL Experimental Center, 2009.

[Alligier 12] R. Alligier, D. Gianazza and N. Durand. *Energy Rate Prediction Using an Equivalent Thrust Setting Profile.* 5th International Conference on Research in Air Transportation (ICRAT 2012), May 22-25, 2012, University of California, Berkeley, USA, 2012.

[Alligier 13] R. Alligier, D. Gianazza and N. Durand. *Ground-based Estimation of the Aircraft Mass, Adaptive vs. Least Squares Method.* ATM 2013, Chicago, USA, June, 2013.

[Anderson 52] T. W. Anderson and D. A. Darling. *Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes.* The Annals of Mathematical Statistics, vol. 23, no. 2, pages 193–212, 1952.

[Aregui 07a] A. Aregui and T. Denoeux. *Consonant Belief Function Induced by a Confidence Set of Pignistic Probabilities.* Khaled Mellouli, editor, Symbolic and Quantitative Approaches to Reasoning with Uncertainty, volume 4724 de *Lecture Notes in Computer Science*, pages 344–355. Springer Berlin Heidelberg, 2007.

[Aregui 07b] A. Aregui and T. Denœux. *Constructing Predictive Belief Functions from Continuous Sample Data Using Confidence Bands.* ISIPTA, pages 11–20, July 2007.

[Aregui 07c] A. Aregui and T. Denoeux. *Fusion of one-class classifiers in the belief function framework.* 10th International Conference on Information Fusion, pages 1–8, 2007.

[Arnold 98] B. C. Arnold and R. M. Shavelle. *Joint Confidence Sets for the Mean and Variance of a Normal Distribution.* The American Statistician, vol. 52, no. 2, pages 133–140, 1998.

[Atkeson 97] C. G. Atkeson, A. W. Moore and Schaal S. *Locally weighted learning.* Artificial Intelligence Review, pages 11–73, 1997.

[Bahadur 66] R. R. Bahadur. *A Note on Quantiles in Large Samples.* The Annals of Mathematical Statistics, vol. 37, no. 3, pages 577–580, 1966.

[Baudrit 06] C. Baudrit and D. Dubois. *Practical representations of incomplete probabilistic knowledge.* Computational Statistics And Data Analysis, vol. 51, no. 1, pages 86–108, November 2006.

[Bayes 63] T. Bayes. *An essay towards solving a problem in the doctrine of chances.* Philosophical Transactions of the Royal Society of London, vol. 53, pages 370–418, 1763.

[Berger 84] J.O. Berger. *The robust Bayesian viewpoint (with discussion).* Robustness of Bayesian Analyses, pages 63–144, 1984.

[Berk 78] R. H. Berk and D. H. Jones. *Relatively Optimal Combinations of Test Statistics.* Scandinavian Journal of Statistics, vol. 5, no. 3, pages 158–162, 1978.

[Berleant 93] D. Berleant. *Automatically verified reasoning with both intervals and probability density functions.* Interval Computations, vol. 2, no. 1993, pages 48–70, 1993.

[Bhattacharya 90] P. K. Bhattacharya and Ashis K. Gangopadhyay. *Kernel and Nearest-Neighbor Estimation of a Conditional Quantile.* Annals of Statistics, vol. 18, no. 3, pages 1400–1415, 1990.

[Billard 00] L. Billard and E. Diday. *Regression Analysis for Interval-Valued Data.* Proceedings of the Seventh Conference of the International Federation of Classification Societies, Springer-Verlag, pages 369–374, 2000.

[Billard 02] L. Billard and E. Diday. *Symbolic regression analysis.* Classification, Clustering and Data Analysis, Proceedings of the 8th Conference of the International Federation of Classification Societies (IFCS'02), Springer, Poland, pages 281–288, 2002.

[Billard 12] L. Billard and E. Diday. Symbolic data analysis: Conceptual statistics and data mining. Wiley Series in Computational Statistics. Wiley, 2012.

[Birnbaum 52] Z. W. Birnbaum. *Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size.* Journal of the American Statistical Association, vol. 47, no. 259, pages 425–441, 1952.

[Bock 00] H.H. Bock and E. Diday. Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, 2000.

[Boole 54] G. Boole. An investigation of the laws of thought: On which are founded the mathematical theories of logic and probabilities. George Boole's collected logical works. Walton and Maberly, 1854.

[Bounhas 13] M. Bounhas, M. Ghasemi Hamed, H. Prade, M. Serrurier and K. Mellouli. *Naive possibilistic classifiers for imprecise or uncertain numerical data.* Fuzzy Sets and Systems, no. 0, pages –, 2013.

[Carroll 88] R.J. Carroll and D. Ruppert. Transformation and weighting in regression. Monographs on Statistics and Applied Probability. Chapman and Hall, 1988.

[Cattaneo 11] M. Cattaneo and A. Wiencierz. *Regression with imprecise data: A robust approach.* ISIPTA11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications, eds. F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, pages 119–128, 2011.

[Cattaneo 12] M. Cattaneo and A. Wiencierz. *Likelihood-based Imprecise Regression.* International Journal of Approximate Reasoning, vol. 53, no. 8, pages 1137–1154, November 2012.

[Celmins 87] A. Celmins. *Least squares model fitting to fuzzy vector data.* Fuzzy Sets and Systems, vol. 22, no. 3, pages 245–269, 1987.

[Center 98] EUROCONTROL Experimental Center. *Study of the Acquisition of Data from Aircraft Operators to Aid Trajectory Prediction Calculation.* Technical report, EUROCONTROL Experimental Center, 1998.

[Chakraborti 00] S. Chakraborti and P. van der Laan. *Precedence Probability and Prediction Intervals.* Journal of the Royal Statistical Society Series D (The Statistician), vol. 49, no. 2, pages 219–228, 2000.

[Chaloulos 10] G. Chaloulos, E. Crück and J. Lygeros. *A simulation based study of subliminal control for air traffic management.* Transportation Research Part C:Emerging Technologies, vol. 18, no. 6, pages 963–974, 2010.

[Chaudhuri 91] P. Chaudhuri. *Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation.* Annals of Statistics, vol. 19, no. 2, pages 760–777, 1991.

[Chen 93] J. Chen and J. Shao. *Iterative Weighted Least Squares Estimators*. Annals of Statistics, vol. 21, no. 2, pages 1071–1092, 1993.

[Cheng 83] R. C. H. Cheng and T. C. Iles. *Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables*. Technometrics, vol. 25, no. 1, pages 77–86, 1983.

[Cheng 01] C. Cheng and E. Stanley Lee. *Fuzzy regression with radial basis function network*. Fuzzy Sets and Systems, vol. 119, no. 2, pages 291–301, 2001.

[Chew 66] V. Chew. *Confidence, Prediction, and Tolerance Regions for the Multivariate Normal Distribution*. Journal of the American Statistical Association, vol. 61, no. 315, pages 605–617, 1966.

[Civanlar 86] M. R Civanlar and H. J. Trussell. *Constructing membership functions using statistical data*. Fuzzy Sets and Systems, vol. 18, pages 1–13, January 1986.

[Cleveland 79] W. S. Cleveland. *Robust Locally Weighted Regression and Smoothing Scatterplots*. Journal of the American Statistical Association, vol. 74, no. 368, pages 829–836, 1979.

[Cleveland 88] W. S. Cleveland and S. J. Devlin. *Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting*. Journal of the American Statistical Association, vol. 83, no. 403, pages 596–610, 1988.

[Consortium 07] SESAR Consortium. *Milestone Deliverable D3: The ATM Target Concept*. Technical report, 2007.

[Coppenbarger 99] R. A. Coppenbarger. *Climb trajectory prediction enhancement using airline flight-planning information*. AIAA Guidance, Navigation, and Control Conference, 1999.

[Cortez 98] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties*. Decision Support Systems, vol. 47, no. 4, pages 547–553, 1998.

[De Brabanter 11] K. De Brabanter, J. De Brabanter, J. A K Suykens and B. De Moor. *Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression*. Neural Networks, IEEE Transactions on, vol. 22, no. 1, pages 110–120, 2011.

[de Campos 94] L. M. de Campos, J. F. Huete and M. Serafin. *Probability intervals: a tool for uncertain reasoning*. International Journal of Uncertainty,

Fuzziness and Knowledge-Based Systems, vol. 2, no. 02, pages 167–196, 1994.

[De Finetti 72] B. De Finetti. Probability, induction and statistics: The art of guessing. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. J. Wiley, 1972.

[de Lima Neto 08] A. E. de Lima Neto and F. A.T. de Carvalho. *Centre and Range method for fitting a linear regression model to symbolic interval data.* Computational Statistics and Data Analysis, vol. 52, no. 3, pages 1500–1515, 2008.

[de Lima Neto 09] A. E. de Lima Neto, G.M. Cordeiro, F.A.T. de Carvalho, U.U. dos Anjos and A.G. da Costa. *Bivariate Generalized Linear Model for Interval-Valued Variables.* Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pages 2226 –2229, june 2009.

[Delgado 87] M. Delgado and S. Moral. *On the concept of possibility-probability consistency.* Fuzzy Sets and Systems, vol. 21, no. 3, pages 311–318, 1987.

[Denoeux 04] T. Denoeux and M.-H. Masson. *EVCLUS: evidential clustering of proximity data.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 34, no. 1, pages 95–109, 2004.

[Denoeux 06] T. Denoeux. *Constructing belief functions from sample data using multinomial confidence regions.* International Journal of Approximate Reasoning, vol. 42, page 2006, 2006.

[Destercke 08] S. Destercke, D. Dubois and E. Chojnacki. *Unifying practical uncertainty representations I: Generalized p-boxes.* International Journal of Approximate Reasoning, vol. 49, no. 3, pages 649–663, 2008.

[Diamond 88] P. Diamond. *Fuzzy least squares.* Information Science, vol. 46, no. 3, pages 141–157, 1988.

[Didier 06] D. Didier. *Possibility theory and statistical reasoning.* Computational Statistics and Data Analysis, vol. 51, pages 47–69, 2006.

[Domingues 10] M. A .O. Domingues, R. M.C.R. de Souza and F. J. A. Cysneiros. *A robust method for linear regression of symbolic interval data.* Pattern Recognition Letters, vol. 31, no. 13, pages 1991–1996, 2010.

[Draper 79] N. R. Draper and R. C. V. Nostrand. *Ridge Regression and James-Stein Estimation: Review and Comments.* Technometrics, vol. 21, no. 4, pages 451–466, 1979.

[Drogoul 09] F. Drogoul, P. Averty and R. Weber. *ERASMUS Strategic Deconfliction to Benefit SESAR.* Proceedings of the 8th USA/Europe Air Traffic Management R&D Seminar, June-July 2009.

[Dubois 80] D. Dubois and H. Prade. Fuzzy sets and systems - Theory and applications. Academic press, New York, 1980.

[Dubois 93a] D. Dubois and H. Prade. *Fuzzy sets and probability : Misunderstandings, bridges and gaps.* IEEE Fuzzy Sys., pages 1059–1068, 1993.

[Dubois 93b] D. Dubois, H. Prade and Sandra Sandri. *On Possibility/Probability Transformations.* IFSA, pages 103–112, 1993.

[Dubois 98] D. Dubois and H. Prade. *Possibility theory:qualitative and quantitative aspects.* D. M. Gabbay and P. Smets, editors, Quantified Representation of Uncertainty and Imprecision, volume 1 de *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, pages 169–226. Kluwer Academic Publishers, 1998. BB.

[Dubois 04] D. Dubois, L. Foulloy, G. Mauris and H. Prade. *Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities.* Reliable Computing, vol. 10, page 2004, 2004.

[Durand 96] N. Durand, J.M. Alliot and J. Noailles. *Automatic aircraft conflict resolution using Genetic Algorithms.* Proceedings of the Symposium on Applied Computing, Philadelphia. ACM, 1996.

[Efron 04] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. *Least Angle Regression.* Annals of Statistics, vol. 32, no. 2, pages 407–451, 2004.

[eng 11] *Engineering Statistics Handbook*, 2011.

[Eubank 93] R. L. Eubank and P. L. Speckman. *Confidence Bands in Nonparametric Regression.* Journal of the American Statistical Association, vol. 88, no. 424, pages 1287–1301, 1993.

[Eubank 99] R.L. Eubank. Nonparametric regression and spline smoothing, second edition. Statistics: A Series of Textbooks and Monogrphs. Marcel Dekker, 1999.

[Fan 92a] J. Fan. *Design-adaptive Nonparametric Regression.* Journal of the American Statistical Association, vol. 87, no. 420, pages 998–1004, 1992.

[Fan 92b] J. Fan and I. Gijbels. *Variable Bandwidth and Local Linear Regression Smoothers*. Annals of Statistics, vol. 20, no. 4, pages 2008–2036, 1992.

[Fan 93] J. Fan. *Local Linear Regression Smoothers and Their Minimax Efficiencies*. Annals of Statistics, vol. 21, no. 1, pages 196–216, 1993.

[Fan 94] J. Fan and J. S. Marron. *Fast Implementations of Nonparametric Curve Estimators*. Journal of Computational and Graphical Statistics, vol. 3, no. 1, pages 35–56, 1994.

[Fan 95] J. Fan, N. E. Heckman and M. P. Wand. *Local polynomial kernel regression for generalized linear models and quasi-likelihood functions*. Journal of the American Statistical Association,90, pages 141–150, 1995.

[Fan 96] J. Fan and I. Gijbels. Local polynomial modelling and its applications: Monographs on statistics and applied probability 66. Monographs on Statistics and Applied Probability, 66. Chapman & Hall, 1996.

[Ferson 03] S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers and K. Sentz. *Constructing Probability Boxes and Dempster-Shafer Structures*. SAND2002-4015, no. January, pages 1–143, 2003.

[Fine 88] T. L. Fine. *Lower probability models for uncertainty and nondeterministic processes*. Journal of Statistical Planning and Inference, vol. 20, no. 3, pages 389–411, 1988.

[Fisher 59] R. A. Fisher. Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh, second edition, 1959.

[Frank 10] A. Frank and A. Asuncion. *UCI Machine Learning Repository*, 2010.

[Frey 08] J. Frey. *Optimal distribution-free confidence bands for a distribution function*. Journal of Statistical Planning and Inference, vol. 138, no. 10, pages 3086–3098, 2008.

[Frey 09] J. Frey, O. Marrero and D. Norton. *Minimum-area confidence sets for a normal distribution*. Journal of Statistical Planning and Inference., vol. 139, pages 1023–1032, 2009.

[Friedman 81] J. H. Friedman and W. Stuetzle. *Projection pursuit regression*. Journal of the American Statistical Association, vol. 76, pages 817–823, 1981.

[Friedman 84] J. Friedman. A variable span smoother. LCS technincal report. Stanford University. Dept. of Statistics. Laboratory for Computational Statistics, 1984.

[Friedman 91] J. H. Friedman. *Multivariate Adaptive Regression Splines.* Annals of Statistics, vol. 19, no. 1, pages 1–67, 1991.

[Gasser 86] T. Gasser, L. Sroka and C. Jennen-Steinmetz. *Residual Variance and Residual Pattern in Nonlinear Regression.* Biometrika, vol. 73, no. 3, pages 625–633, 1986.

[Gasser 89] T. Gasser and A. Kneip. *Discussion: Linear Smoothers and Additive Models.* Annals of Statistics, vol. 17, no. 2, pages 532–535, 1989.

[Ghasemi Hamed 12a] M. Ghasemi Hamed, M. Serrurier and D. Durand. *Possibilistic KNN regression using tolerance intervals.* IPMU 2012, Catania, Italy, volume 299 de *Communications in Computer and Information Science*. Springer, July 2012.

[Ghasemi Hamed 12b] M. Ghasemi Hamed, M. Serrurier and D. Durand. *Representing Uncertainty by Possibility Distributions Encoding Confidence Bands, Tolerance and Prediction Intervals.* SUM, pages 233–246, 2012.

[Ghasemi Hamed 12c] M. Ghasemi Hamed, M. Serrurier and D. Durand. *Simultaneous Interval Regression for K-Nearest Neighbor.* Australasian Conference on Artificial Intelligence, pages 602–613, 2012.

[Ghasemi Hamed 13] M. Ghasemi Hamed, D. Gianazza, M. Serrurier and N. Durand. *Statistical prediction of aircraft trajectory: regression methods vs point-mass model.* ATM 2013, Chicago, USA, June, 2013.

[Gibbons 94] R.D. Gibbons. Statistical methods for groundwater monitoring. A Wiley-Interscience publication. Wiley, 1994.

[Gibbons 01] R.D. Gibbons and D.E. Coleman. Statistical methods for detection and quantification of environmental contamination. Wiley, 2001.

[Godfrey 09] LG Godfrey. Bootstrap tests for regression models. Palgrave Macmillan, 2009.

[Gong 04] C. Gong and D. McNally. *A methodology for automated trajectory prediction analysis.* AIAA Guidance, Navigation, and Control Conference and Exhibit, 2004.

[Good 62] I. J. Good. *Subjective Probability as the Measure of a Non-measurable Set.* 1962.

[Good 65] I.J. Good. The estimation of probabilities: an essay on modern bayesian methods. Research monograph. M.I.T. Press, 1965.

[Grize 87] Y. L. Grize and T. L. Fine. *Continuous Lower Probability-Based Models for Stationary Processes with Bounded and Divergent Time Averages.* The Annals of Probability, vol. 15, no. 2, pages 783–803, 1987.

[Hahn 69] G.J. Hahn. *Factors for Calculating Two-Sided Prediction Intervals for Samples from a Normal Distribution.* Journal of the American Statistical Association, vol. 64, no. 327, pages 878–888, 1969.

[Hahn 91] G. J. Hahn and W. Q. Meeker. Statistical intervals: A guide for practitioners. John Wiley and Sons, 1991.

[Hanson 63] D. L. Hanson and D. B. Owen. *Distribution-Free Tolerance Limits Elimination of the Requirement That Cumulative Distribution Functions Be Continuous.* Technometrics, vol. 5, no. 4, pages 518–522, 1963.

[Härdle 90] W. Härdle. Applied nonparametric regression. Econometric Society Monographs (No. 19). Cambridge University Press, 1990.

[Hardle 91] W. Hardle and J. S. Marron. *Bootstrap Simultaneous Error Bars for Nonparametric Regression.* Annals of Statistics, vol. 19, no. 2, pages 778–796, 1991.

[Hastie 86] T. Hastie and R. Tibshirani. *Generalized additive models.* Statistical Science, vol. 1, pages 297–310, 1986.

[Hastie 90] T.J. Hastie and R.J. Tibshirani. Generalized additive models. Monographs on Statistics and Applied Probability. Chapman & Hall, 1990.

[Hastie 93] T. Hastie and C. Loader. *Local Regression: Automatic Kernel Carpentry.* Statistical Science, vol. 8, no. 2, pages 120–129, 1993.

[He 96] X. He and Q. Shao. *A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs.* Annals of Statistics, vol. 24, no. 6, pages 2608–2630, 1996.

[He 02] X. He and F. Hu. *Markov Chain Marginal Bootstrap.* Journal of the American Statistical Association, vol. 97, no. 459, pages 783–795, 2002.

[Hong 03] D. H. Hong and C. Hwang. *Support vector fuzzy regression machines.* Fuzzy Sets and Systems, vol. 138, no. 2, pages 271–281, 2003.

[Hong 05] D. H. Hong and C. Hwang. *Interval regression analysis using quadratic loss support vector machine.* Trans. Fuz Sys., vol. 13, no. 2, pages 229–237, April 2005.

[Howe 69] W. G. Howe. *Two-Sided Tolerance Limits for Normal Populations, Some Improvements.* Journal of the American Statistical Association, vol. 64, no. 326, pages 610–620, 1969.

[Huang 98] L. Huang, B. L. Zhang and Q. Huang. *Robust interval regression analysis using neural networks.* Fuzzy Sets and Systems, vol. 97, no. 3, pages 337–347, 1998.

[Huber 73] P. J. Huber and V. Strassen. *Minimax Tests and the Neyman-Pearson Lemma for Capacities.* Annals of Statistics, vol. 1, no. 2, pages 251–263, 1973.

[Huber 09] P.J. Huber and E.M. Ronchetti. Robust statistics. Wiley Series in Probability and Statistics. Wiley, 2009.

[Ishibuchi 90] H. Ishibuchi and H. Tanaka. *Several formulations of interval regression analysis.* Proceedings Sino-Japan Joint Meeting on Fuzzy Sets and Systems, Beijing,China, 1990.

[Ishibuchi 92] H. Ishibuchi and H. Tanaka. *Fuzzy regression analysis using neural networks.* Fuzzy Sets and Systems, vol. 50, no. 3, pages 257–265, 1992.

[Ishibuchi 93] H. Ishibuchi, H. Tanaka and H. Okada. *An architecture of neural networks with interval weights and its application to fuzzy regression analysis.* Fuzzy Sets and Systems, vol. 57, no. 1, pages 27–39, 1993.

[Iwasaki 05] M. Iwasaki and H. Tsubaki. *A bivariate generalized linear model with an application to meteorological data analysis.* Statistical Methodology, vol. 2, no. 3, pages 175–190, 2005.

[Jeng 03] J. T. Jeng, Chuang C. C. and S. F. Su. *Support vector interval regression networks for interval regression analysis.* Fuzzy Sets and Systems, vol. 138, no. 2, pages 283–300, 2003.

[John 63] S. John. *A Tolerance Region for Multivariate Normal Distributions.* Sankhyā: The Indian Journal of Statistics, Series A, vol. 25, no. 4, pages 363–368, 1963.

[Kanofsky 72] P. Kanofsky and R. Srinivasan. *An Approach to the Construction of Parametric Confidence Bands on Cumulative Distribution Functions.* Biometrika, vol. 59, no. 3, pages 623–631, 1972.

[Klir 90] G.J. Klir. *A principle of uncertainty and information invariance.* International Journal of General Systems, vol. 17, no. 23, pages 249–275, 1990.

[Kocherginsky 05] M. Kocherginsky, X. He and Y. Mu. *Practical Confidence Intervals for Regression Quantiles.* Journal of Computational and Graphical Statistics, vol. 14, no. 1, pages 41–55, 2005.

[Koenker 78] R. Koenker and G. Bassett. *Regression Quantiles.* Econometrica, vol. 46, no. 1, pages 33–50, 1978.

[Koenker 94a] R. Koenker. *Confidence intervals for regression quantiles.* Asymptotic statistics, pages 349–359. Springer, 1994.

[Koenker 94b] R. Koenker, P. Ng and S. Portnoy. *Quantile smoothing splines.* Biometrika, vol. 81, no. 4, pages 673–680, 1994.

[Koenker 96] R. Koenker and J. B. Park. *An interior point algorithm for nonlinear quantile regression.* Journal of Econometrics, vol. 71, pages 265–283, 1996.

[Koenker 01] R. Koenker and K. Hallock. *Quantile Regression: An Introduction.* Journal of Economic Perspectives, vol. 15, no. 4, pages 43–56, 2001.

[Koenker 05] R. Koenker. Quantile regression. Econometric Society Monographs. Cambridge University Press, 2005.

[Konijn 87] H. S. Konijn. *Distribution-Free and Other Prediction Intervals.* The American Statistician, vol. 41, no. 1, pages 11–15, 1987.

[Krishnamoorthy 99] K. Krishnamoorthy and Thomas Mathew. *Comparison of Approximation Methods for Computing Tolerance Factors for a Multivariate Normal Population.* Technometrics, vol. 41, no. 3, pages 234–249, 1999.

[Krishnamoorthy 09] K. Krishnamoorthy and T. Mathew. Statistical tolerance regions: Theory, applications, and computation. Wiley Series in Probability and Statistics. Wiley, 2009.

[Krishnamoorthya 06] K. Krishnamoorthya and S. Mondala. *Improved Tolerance Factors for Multivariate Normal Distributions.* Communications in Statistics - Simulation and Computation, vol. 35, no. 2, pages 461–478, 2006.

[Laplace 12] P.S. Laplace. Théorie analytique des probabilités. Courcier, Paris, 1812.

[Laplace 14] P.S. Laplace. Essai philosophique sur les probabilités;. Mme. Ve. Courcier, 1814.

[Li 07] Q. Li and J.S. Racine. Nonparametric econometrics: theory and practice. Princeton University Press, 2007.

[Lieberman 63] G. J. Lieberman and Jr. Miller R. G. *Simultaneous Tolerance Intervals in Regression.* Biometrika, vol. 50, no. 1/2, pages 155–168, 1963.

[Lymperopoulos 06] I. Lymperopoulos, J. Lygeros and A. Lecchini Visintini. *Model Based Aircraft Trajectory Prediction during Takeoff.* AIAA Guidance, Navigation and Control Conference and Exhibit, Keystone, Colorado, August 2006.

[Masson 04] M.-H. Masson and T. Denoeux. *Clustering interval-valued proximity data using belief functions.* Pattern Recognition Letters, vol. 25, no. 2, pages 163–171, 2004.

[Masson 06] M. Masson and T. Denoeux. *Inferring a possibility distribution from empirical data.* Fuzzy Sets and Systems, vol. 157, pages 319–340, February 2006.

[Mee 91] R. W. Mee, K. R. Eberhardt and C. P. Reeve. *Calibration and Simultaneous Tolerance Intervals for Regression.* Technometrics, vol. 33, no. 2, pages 211–219, 1991.

[Mendenhall 06] W. Mendenhall and T. Sincich. Statistics for engineering and the sciences (5th edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[Mercier 08] D. Mercier, Quost B. and T. Denoeux. *Refined modeling of sensor reliability in the belief function framework using contextual discounting.* Information Fusion, vol. 9, no. 2, pages 246–258, 2008.

[Miller 91] R. G. Miller. Simultaneous Statistical Inference. Springer-Verlag, New York, 1991.

[Murphy 48] R. B. Murphy. *Non-Parametric Tolerance Limits.* The Annals of Mathematical Statistics, vol. 19, no. 4, pages 581–589, 1948.

[Musialek 10] B. Musialek, C. F. Munafo, H. Ryan and M. Paglione. *Literature Survey of Trajectory Predictor Technology.* Technical report DOT/FAA/TC-TN11/1, Federal Aviation Administration William J. Hughes Technical Center, Atlantic City, November 2010.

[Nadaraya 64] E. A. Nadaraya. *On estimating regression.* Theory of Probability and its Applications, vol. 9, pages 141–142, 1964.

[Nuic 09] A. Nuic. *User Manual for Base of Aircarft DATA (BADA) Rev.3.7.* Technical report, EUROCONTROL, 2009.

[Owen 95] A. B. Owen. *Nonparametric Likelihood Confidence Bands for a Distribution Function.* Journal of the American Statistical Association, vol. 90, no. 430, pages 516–521, 1995.

[Paulson 43] E. Paulson. *A Note on Tolerance Limits.* The Annals of Mathematical Statistics, vol. 14, no. 1, pages 90–93, 1943.

[Peters 94] G. Peters. *Fuzzy linear regression with fuzzy intervals.* Fuzzy Sets and Systems, vol. 63, no. 1, pages 45–55, 1994.

[Petit-Renaud 04] S. Petit-Renaud and T. Denoeux. *Nonparametric regression analysis of uncertain and imprecise data using belief functions.* International Journal of Approximate Reasoning, vol. 35, no. 1, pages 1–28, 2004.

[Prats 10] X. Prats, V. Puig, J. Quevedo and F. Nejjari. *Multi-objective optimisation for aircraft departure trajectories minimising noise annoyance.* Transportation Research Part C, vol. 18, no. 6, pages 975–989, 2010.

[Quost 07] B. Quost, T. Denoeux and M. H. Masson. *Pairwise classifier combination using belief functions.* Pattern Recognition Letters, vol. 28, no. 5, pages 644–653, 2007.

[Rao 99] C. R. Rao and H. Toutenburg. Linear Models: Least Squares and Alternatives (Springer Series in Statistics). Springer, July 1999.

[Raufaste 03] Eric Raufaste, Rui da Silva Neves and Claudette Mariné. *Testing the descriptive validity of possibility theory in human judgments of uncertainty.* Artificial Intelligence, vol. 148, no. 1, pages 197–218, 2003.

[Romanelli 09] J. Romanelli, C. Santiago, M. Paglione and A. Schwartz. *Climb trajectory prediction software validation for decision support tools and simulation models.* International Test and Evaluation Association, 2009.

[Ruppert 94] D. Ruppert and M. P. Wand. *Multivariate Locally Weighted Least Squares Regression.* Annals of Statistics, vol. 22, no. 3, pages 1346–1370, 1994.

[Savage 72] L.J. Savage. The foundations of statistics. Dover Books on Mathematics Series. DOVER PUBN Incorporated, 1972.

[Scheffé 59] H. Scheffé. The analysis of variance. A Wiley publication in mathematical statistics. Wiley, 1959.

[Seifert 94] Burkhardt Seifert, Michael Brockmann, Joachim Engel and Theo Gasser. *Fast Algorithms for Nonparametric Curve Estimation.* Journal of Computational and Graphical Statistics, vol. 3, no. 2, pages 192–213, 1994.

[Serrurier 07] M. Serrurier and H. Prade. *A general framework for imprecise regression.* Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International, pages 1–6, 2007.

[Shafer 76] G. Shafer. A mathematical theory of evidence. Limited paperback editions. Princeton University Press, 1976.

[Silverman 85] B. W. Silverman. *Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 47, no. 1, pages 1–52, 1985.

[Silverman 86] B. W. Silverman. Density estimation for statistics and data analysis. Chapman and Hall, London; New York, 1986.

[Slotani 64] M. Slotani. *Tolerance regions for a multivariate normal population.* Annals of the Institute of Statistical Mathematics, vol. 16, pages 135–153, 1964.

[Smets 90a] P. Smets. *Fusion of one-class classifiers in the belief function framework.* NAFIPS-90, pages 215–218, 1990.

[Smets 90b] P. Smets. *The Transferable Belief Model and Other Interpretations of Dempster-Shafer's Model.* pages 326–333, 1990.

[Smets 94] P. Smets and R. Kennes. *The transferable belief model.* Artificial Intelligence, vol. 66, no. 2, pages 191–234, 1994.

[Smets 13] P. Smets. *Practical Uses of Belief Functions.* CoRR, vol. abs/1301.6741, 2013.

[Stone 77a] C. J. Stone. *Consistent Nonparametric Regression.* Annals of Statistics, vol. 5, no. 4, pages 595–620, 1977.

[Stone 77b] C. J. Stone. *Consistent Nonparametric Regression.* Annals of Statistics, vol. 5, no. 4, pages 595–620, 1977.

[Su 13] Zhi-gang Su and Pei-hong Wang. *Regression analysis of belief functions on interval-valued variables: comparative studies.* Soft Computing, pages 1–20, 2013.

[Sun 94] J. Sun and C. R. Loader. *Simultaneous Confidence Bands for Linear Regression and Smoothing.* Annals of Statistics, vol. 22, no. 3, pages 1328–1345, 1994.

[Swenson 06] H. Swenson, R. Barhydt and M. Landis. *Next Generation Air Transportation System (NGATS) Air Traffic Management (ATM)-Airspace Project.* Technical report, National Aeronautics and Space Administration, 2006.

[T. Kinoshita 06] F. Imado T. Kinoshita. *The Application of an UAV Flight Simulator - The Development of a New Point Mass Model for an Aircraft.* SICE-ICASE International Joint Conference Conference, 2006.

[Takeuchi 06] I. Takeuchi, Q. V. Le, T. D. Sears and A. J. Smola. *Nonparametric Quantile Estimation.* Journal of Machine Learning Research, vol. 7, pages 1231–1264, December 2006.

[Tanaka 87] H. Tanaka. *Fuzzy data analysis by possibilistic linear models.* Fuzzy Sets and Systems, vol. 24, no. 3, pages 363–376, 1987.

[Tastambekov 14] K. Tastambekov, S. Puechmorel, D. Delahaye and C. Rabut. *Aircraft trajectory forecasting using local functional regression in Sobolev space.* Transportation Research Part C: Emerging Technologies, vol. 39, no. 0, pages 1 – 22, 2014.

[Technology 91] R.M.C.A.P.M.I.D.U. Technology and K.S.F.A.P.M.I.D.U. Technology. Experts in uncertainty : Opinion and subjective probability in science: Opinion and subjective probability in science. Environmental ethics and science policy. Oxford University Press, USA, 1991.

[Tibshirani 96] R. Tibshirani. *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society Series B (Methodological), vol. 58, no. 1, pages 267–288, 1996.

[Tukey 47] J. W. Tukey. *Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions-The Continuous Case.* The Annals of Mathematical Statistics, vol. 18, no. 4, pages 529–539, 1947.

[Tukey 48] J. W. Tukey. *Nonparametric Estimation, III. Statistically Equivalent Blocks and Multivariate Tolerance Regions-The Discontinuous Case.* The Annals of Mathematical Statistics, vol. 19, no. 1, pages 30–39, 1948.

[Vivona 10] R. A. Vivona, M. M. Paglione, K. T. Cate and G. Enea. *Definition and demonstration of a methodology for validating aircraft trajectory predictors*. AIAA Guidance, Navigation, and Control Conference, 2010.

[Wahba 90] G. Wahba. Spline models for observational data, volume 59 de *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Wald 42] A. Wald. *Setting of Tolerance Limits When the Sample is Large*. The Annals of Mathematical Statistics, vol. 13, no. 4, pages 389–399, 1942.

[Wald 43] A. Wald. *An Extension of Wilks' Method for Setting Tolerance Limits*. The Annals of Mathematical Statistics, vol. 14, no. 1, pages 45–55, 1943.

[Wald 46] A. Wald and J. Wolfowitz. *Tolerance Limits for a Normal Distribution*. The Annals of Mathematical Statistics, vol. 17, no. 2, pages 208–215, 1946.

[Walley 82] P. Walley and T. L. Fine. *Towards a Frequentist Theory of Upper and Lower Probability*. Annals of Statistics, vol. 10, no. 3, pages 741–761, 1982.

[Walley 91] P. Walley. Statistical reasoning with imprecise probabilities. Chapman and Hall, 1991.

[Wallis 51] W. A. Wallis. *Tolerance Intervals for Linear Regression*. Proceedings Second Berkeley Symposium on Mathematical Statistics and Probability, pages 43–51. Univ. of Calif. Press, 1951.

[Watson 64] G. S. Watson. *Smooth regression analysis*. Sankhyā: The Indian Journal of Statistics, Series A, vol. 26, pages 359–372, 1964.

[Wiencierz 13] A. Wiencierz and M. Cattaneo. *An Exact Algorithm for Likelihood-Based Imprecise Regression in the Caseof Simple Linear Regression with Interval Data*. vol. 190, pages 293–301, 2013.

[Wilks 41] S. S. Wilks. *Determination of Sample Sizes for Setting Tolerance Limits*. The Annals of Mathematical Statistics, vol. 12, no. 1, pages 91–96, 1941.

[Williamson 89] R. C. Williamson. *Probabilistic Arithmetic*. Technical report, University of Queensland, Austrialia, 1989.

[Williamson 90] R. C Williamson and T. Downs. *Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds.* International Journal of Approximate Reasoning, vol. 4, no. 2, pages 89–158, 1990.

[Wilson 67] A. L. Wilson. *An Approach to Simultaneous Tolerance Intervals in Regression.* The Annals of Mathematical Statistics, vol. 38, no. 5, pages 1536–1540, 1967.

[Working 29] H. Working and H. Hotelling. *Applications of the Theory of Error to the Interpretation of Trends.* Journal of the American Statistical Association, vol. 24, no. 165, pages 73–85, 1929.

[Yeh 98] I.-C. Yeh. *Modeling of strength of high-performance concrete using artificial neural networks.* Cement and Concrete Research, vol. 28, no. 12, pages 1797–1808, 1998.

[Yeh 07] I-Cheng Yeh. *Modeling slump flow of concrete using second-order regressions and artificial neural networks.* Cement and Concrete Composites, vol. 29, no. 6, pages 474–480, 2007.

[Yu 04] K. Yu and M.C. Jones. *Likelihood-based local linear estimation of the conditional variance function.* Journal of the American Statistical Association, vol. 99, no. 465, pages 139–144, March 2004.

[Zadeh 78] L.A. Zadeh. *Fuzzy sets as a basis for a theory of possibility.* Fuzzy Sets and Systems, vol. 1, no. 1, pages 3–28, 1978.

Planification de trajectoires pour l'optimisation du trafic aérien

Résumé : Le trafic aérien en Europe représente environ 30 000 vols quotidiens actuellement. Selon les prévisions de l'organisme Eurocontrol, ce trafic devrait croître de 70 % d'ici l'année 2020 pour atteindre 50 000 vols quotidiens. L'espace aérien, découpé en zones géographiques appelées secteurs de contrôle, atteindra bientôt son niveau de saturation vis-à-vis des méthodes actuelles de planification et de contrôle.

Afin d'augmenter la quantité de trafic que peut absorber le système, il est nécessaire de diminuer la charge de travail des contrôleurs aériens en les aidant dans leur tâche de séparation des avions. En se fondant sur les demandes de plans de vol des compagnies aériennes, nous proposons une méthode de planification des trajectoires en 4D permettant de présenter au contrôleur un trafic dont la plupart des conflits auront été évités en avance.

Cette planification s'établit en deux étapes successives, ayant chacune un unique degré de liberté : une allocation de niveaux de vol permettant la résolution des conflits en croisière puis une allocation d'heures de décollage permettant de résoudre les conflits restants. Nous présentons des modèles pour ces deux problèmes d'optimisation fortement combinatoires, que nous résolvons en utilisant la programmation par contraintes ou les algorithmes évolutionnaires, ainsi que des techniques permettant de prendre en compte des incertitudes sur les heures de décollage ou le suivi de trajectoire.

Les simulations conduites sur l'espace aérien français mènent à des situations où tous les conflits sont évités, avec des retards alloués de l'ordre d'une minute en moyenne (80 à 90 minutes pour le vol le plus retardé) et un écart par rapport à l'altitude optimale limité à un niveau de vol pour la quasi totalité des vols. La prise en compte d'incertitudes de manière statique dégrade fortement ces solutions peu robustes, mais nous proposons un modèle dynamique utilisant une fenêtre glissante susceptible de prendre en compte des incertitudes de quelques minutes avec un impact réduit sur le coût de l'allocation.

Mots clés : gestion du trafic aérien, allocation de créneaux de décollage, affectation de niveaux de vol, programmation par contraintes, algorithme évolutionnaire

---

Trajectory planning for air traffic optimization

Abstract : Air traffic in Europe represents about 30,000 flights each day and forecasts from Eurocontrol predict a growth of 70 % by 2020 (50,000 flights per day). The airspace, made up of numerous control sectors, will soon be saturated given the current planification and control methods.

In order to make the system able to cope with the predicted traffic growth, the air traffic controllers workload has to be reduced by automated systems that help them handle the aircraft separation task. Based on the traffic demand by airlines, this study proposes a new planning method for 4D trajectories that provides conflict-free traffic to the controller.

This planning method consists of two successive steps, each handling a unique flight parameter : a flight level allocation phase followed by a ground holding scheme. We present constraint programming models and an evolutionary algorithm to solve these large scale combinatorial optimization problems, as well as techniques for improving the robustness of the model by handling uncertainties of takeoff times and trajectory prediction.

Simulations carried out over the French airspace successfully solved all conflicts, with a mean of one minute allocated delay (80 to 90 minutes for the most delayed flight) and a discrepancy from optimal altitude of one flight level for most of the flights. Handling uncertainties with a static method leads to a dramatic increase in the cost of the previous non-robust solutions. However, we propose a dynamic model to deal with this matter, based on a sliding time horizon, which is likely to be able to cope with a few minutes of uncertainty with reasonable impact on the cost of the solutions.

Keywords : air traffic management, ground holding, flight level allocation, constraint programming, evolutionary algorithm